RODOLFO BOTTO DE BARROS GARCIA

SELEÇÃO DE ATRIBUTOS USANDO CRITÉRIOS MULTIOBJETIVO BASEADOS EM ENXAME PARA SELEÇÃO DE GENES EM MICROARRANJOS (SAMO-ESMA)

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

CURITIBA

2011

RODOLFO BOTTO DE BARROS GARCIA

SELEÇÃO DE ATRIBUTOS USANDO CRITÉRIOS MULTIOBJETIVO BASEADOS EM ENXAME PARA SELEÇÃO DE GENES EM MICROARRANJOS (SAMO-ESMA)

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: *Descoberta do Conhecimento e Aprendizagem de Máquina*

Orientador: Prof. Dr. Júlio Cesar Nievola
Co-orientador: Prof. Dr. Emerson Cabrera Paraiso

CURITIBA

2011

Garcia, Rodolfo Botto de Barros
Seleção de Atributos Usando Critérios Multiobjetivo Baseados em Enxame Para Seleção de Genes em Microarranjos (SAMO-ESMA). Curitiba, 2011. 120p.

Dissertação(Mestrado) – Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática.

1. Mineração de Dados 2. Seleção de Atributos 3. Otimização Multiobjetivo Baseada em Enxame 4. Microarranjo. I.Pontifícia Universidade Católica do Paraná. II. Centro de Ciências Exatas e de Tecnologia. III. Programa de Pós-Graduação em Informática.

Esta página deve ser reservada à ata de defesa e termo de aprovação que serão fornecidos pela secretaria após a defesa da dissertação e efetuadas as correções solicitadas.

Este trabalho é dedicado aos meus pais e irmão por todo o esforço e incentivo que foi dado a mim em todos os momentos de minha vida.

Agradecimentos

Primeiramente a Deus, por estar abençoando meu caminho e iluminando meus pensamentos.

Aos meus pais e ao meu irmão, por todo incentivo e apoio para que eu chegasse até aqui.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, o CNPq, pelo suporte financeiro a este trabalho.

Ao meu orientador Júlio Cesar Nievola e ao meu co-orientador Emerson Cabrera Paraiso, pela confiança em mim depositada, pelos esclarecimentos científicos e por toda aprendizagem ao longo do mestrado.

Aos colegas que pude ter a satisfação de compartilhar este caminho, especialmente ao Jackson Mallmann e ao Lucas Galete.

E a todos os amigos que fiz em Curitiba, por fazer me sentir em casa.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Pesquisas envolvendo o Projeto Genoma Humano têm avançado bastante em relação ao mapeamento genético. Por conta disso, a quantidade de informações armazenadas em bases de dados gênicas tem aumentado muito rapidamente e análises manuais feitas por cientistas demoram até anos para serem completadas.

A análise das expressões gênicas é de extrema importância para diagnosticar doenças e identificar o estado de determinados genes em ciclos específicos. A técnica de microarranjo é responsável pela extração de informação sobre a expressão de grande quantidade de genes em relação a uma determinada característica. O problema é que poucos são os genes que fazem parte da ativação ou inibição dessa característica. A presença de atributos que não influenciam no resultado tornam a análise mais complexa e devem ser removidos.

Por isso, o uso de métodos computacionais cujo objetivo é selecionar os genes mais relevantes para a característica proposta no microarranjo se torna indispensável.

Com a existência de vários métodos de seleção de atributos, é necessária a utilização de vários critérios para garantir a escolha pelo melhor método e, assim, caracterizar um problema de otimização multiobjetivo.

Este trabalho propõe um método que utiliza índices obtidos do agrupamento das instâncias de bases de dados de expressões gênicas para avaliar os métodos de seleção por meio de um método de otimização multiobjetivo eficiente, rápido e baseado em enxames chamado MOPSO (*Multi-Objective Particle Swarm Optimization*).

O resultados dos experimentos mostram que o método proposto é capaz de avaliar seleções de atributos, escolhendo os melhores métodos, e de mostrar que a redução da dimensionalidade das bases de dados melhoram as análises, além de poder revelar os genes mais relevantes para uma determinada característica obtidos pela técnica do microarranjo.

Palavras-Chave: Seleção de Atributos, Agrupamento, Problemas multiobjetivo, microarranjo.

Abstract

Researches involving the Human Genome Project have significantly advances in genetic mapping. For this reason, the amount of information stored in genetics datasets has increased very rapidly and scientists manual analyses take years to be completed.

Analysis of gene expression is extremely important to diagnoses diseases and identify the status of certain genes in specific cycles. The microarray technique is responsible for extracting information about the expression of a large number of genes related to a particular characteristic. The problem is that there are few genes that are present in the activation or inhibition process of these characteristic. The presence of these features that do not influence in the results turn the analysis more complex and have to be removed.

Therefore, is necessary the use of computational methods which aim to select relevant genes to the characteristic proposed in the microarray.

With the existence of various feature selection methods, is necessary use several criteria to ensure the choice for the best method and, thus characterize a multi-objective optimization problem.

This work proposes a method that uses indexes from clustering of datasets instances to evaluate selection methods by a efficient, fast and swarm based multi-objective optimization method called MOPSO (Multi-Objective Particle Swarm Optimization).

The results of experiments show that the proposed method is able to evaluate features selection methods, choosing the best ones and prove that the dimensionality reduction of datasets improves the analyses, also reveals relevant genes to datasets obtained by the microarray technique for one specific characteristic.

Keywords: Feature selection, Clustering, Multi-objective problems, microarray.

Capítulo 1

Introdução

Pesquisas envolvendo o Projeto Genoma Humano têm avançado bastante em relação ao mapeamento genético. Por conta disso, a quantidade de informações armazenadas em bases de dados gênicos tem aumentado muito rapidamente e análises manuais feitas por cientistas podem demorar até anos para terem resultados, necessitando de ajuda computacional para agilizar esse trabalho [Quackenbush, 2001].

O mapeamento dos genes possibilita saber onde eles estão dispostos no Genoma mas não informa o que eles fazem. Dentro de uma célula, nem todos os genes encontram-se ativos e é justamente a diferenciação do conjunto de genes expressos que determina a função biológica celular, assim como o tipo de material que a célula é composta [Brown & Botstein, 1999]. O estudo das expressões gênicas pode esclarecer qual o papel de um determinado gene e quais deles estão relacionados a uma determinada disfunção, já que a má formação ou variação da sequência gênica podem modificar os genes que deverão ser ativados, prejudicando funções do corpo e resultando em doenças [NIH, 2001].

O campo da computação em que são desenvolvidos softwares para manipulação de grandes quantidades de dados da área biológica é chamado de Bioinformática [NIH, 2001]. A Bioinformática oferece à análise gênica, além da alta velocidade de processamento dos computadores, a capacidade de que vários genes possam ser analisados de forma simultânea e armazenados em bases de dados para futuros experimentos [Au et al., 2005].

Análises computacionais mais complexas podem revelar associações entre os genes de um experimento, como por exemplo [Jain, Murty & Flynn, 1999]:

- Se um gene só trabalha mediante a expressão de outro;
- Se vários genes trabalham juntos para executar a mesma função;

- Se alguns genes se mantêm inalterados enquanto outros se expressam;
- Se a expressão de alguns genes é responsável pela inibição de outros;
- Inferir que genes com função desconhecida exercem a mesma função que aqueles que apresentaram comportamento de expressão semelhante.

Uma técnica que explora a análise de expressões gênicas de forma compreensiva é o microarranjo (Figura 1.1). Ele consiste em um chip contendo até milhares de genes que foram obtidos do Genoma mapeado [Quackenbush, 2001].

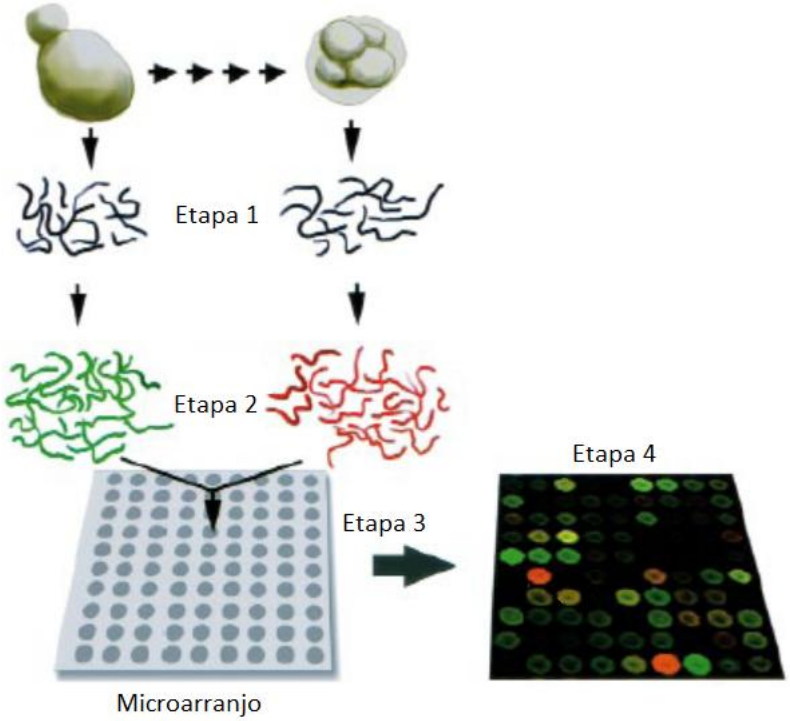

Figura 1.1: Microarranjo de DNA baseado em [Brown & Botstein, 1999]

Na etapa 1, para cada experimento são extraídas amostras gênicas de células sob diferentes condições, como por exemplo a presença de doenças. Essas amostras são compostas por sequências de genes complementares aos genes que já estão no microarranjo que, ao reagirem na etapa 3, determinarão o nível de expressão dessas amostras. Para identificar a origem das sequências, diferentes corantes são usados na etapa 2 e a intensidade de fluorescência equivale ao nível de expressão de cada gene. Um scanner interpreta a expressão na etapa 4 e normaliza os valores de forma numérica para que sua análise seja feita por meio de modelos matemáticos.

O nível de expressão também identifica quais genes estão ativos e inativos em um determinado momento da vida da célula, desvendando aqueles que são responsáveis pela aparição de uma determinada doença, assim como os mais relevantes [NIH, 2001]. Essa técnica tem sido utilizada para buscar os possíveis genes envolvidos em doenças como o câncer, Alzheimer, Parkinson e diabetes [D'haesleer, 2005].

Apesar de extrair informações sobre a expressão de milhares de genes, o problema do microarranjo é que em um único experimento poucas amostras são obtidas e o custo para realizar um experimento é alto, impossibilitando a geração de mais amostras para auxiliar na comparação das expressões gênicas [Yu & Liu, 2004]. Além disso, a expressão de grande quantidade de genes dificulta a análise do microarranjo, pois se sabe que poucos são os genes que fazem parte do processo de ativação ou inibição de uma característica.

Devido a este acúmulo de informação, técnicas de análises de dados e algoritmos de extração de informações significativas, ou de mineração de dados, têm se revelado ferramentas valiosas e de auxílio aos cientistas na análise das inter-relações que permitem a identificação de padrões entre os milhares de genes viabilizando a elucidação de questões biológicas como, por exemplo, as funções que eles desempenham no organismo e os processos biológicos em que os mesmos estão envolvidos [Zhao & Karypis, 2003].

O objetivo deste trabalho é de empregar uma técnica de análise que visa reduzir a dimensionalidade das bases de dados de expressões gênicas obtidas pelo microarranjo, a fim de facilitar sua análise, diminuir o tempo gasto e melhorar os resultados obtidos. Essa técnica, chamada seleção de atributos, busca selecionar os genes mais relevantes considerando seus níveis de expressão nas amostras armazenadas nas bases de dados.

Por existirem diferentes abordagens de seleção de atributos e não haver um consenso sobre qual é a melhor, é necessário aplicar diversas técnicas e comparar seus resultados. A qualidade de um método de seleção pode ser avaliada, de muitas formas, através do agrupamento das amostras que possuem comportamentos similares, verificando se elas pertencem às pessoas sob as mesmas condições. Por sua vez, um agrupamento é avaliado pela otimização dos chamados índices de validação de grupos que, semelhante à técnica de seleção, possuem várias abordagens. Diferentes abordagens são usadas neste trabalho, a fim de se realizar uma avaliação com maior precisão do agrupamento, aproximando-se da precisão caso houvesse uma avaliação manual. O conjunto dos valores desses índices formam soluções que representam cada método de seleção de atributos.

A otimização de vários índices, ou critérios, caracteriza o chamado problema de otimização multiobjetivo. Neste trabalho, o método de otimização multiobjetivo utilizado é baseado em enxames, o qual permite buscas dentro de um espaço de soluções muito grande, ou seja, é possível usá-lo para analisar soluções geradas por vários métodos de seleção de atributos e escolher aquelas que possuam os genes mais relevantes para uma determinada característica.

1.1. Objetivos

Este trabalho tem como objetivo principal avaliar o desempenho de diferentes algoritmos de seleção de atributos, usando critérios variados obtidos na tarefa de agrupamento de dados de microarranjos, combinados através da técnica multiobjetivo baseada em enxames.

Os objetivos específicos englobados são:

a) Estudo das características de um conjunto de bases de dados de microarranjos disponíveis publicamente;

b) Utilização e testes de algoritmos de seleção de atributos baseados na abordagem de filtro e pelos diferentes tipos de busca pelo espaço de atributos (e.g. sequenciais, completas e aleatórias);

c) Implementação e utilização de um conjunto de algoritmos de agrupamento pertencentes a diferentes paradigmas (e.g. hierárquicos e particionais), que formam diferentes tipos de grupos (e.g. baseados em centro, baseados em densidade, baseado em propriedade);

d) Implementação e testes de critérios que avaliam a qualidade dos agrupamentos de dados em bioinformática pertencentes a diferentes abordagens (e.g. critérios internos, externos e relativos);

e) Implementação da técnica multiobjetivo baseada em enxames para combinação dos diversos índices de avaliação;

f) Avaliação do desempenho de cada um dos métodos de seleção de atributos e comparação entre os mesmos.

1.2. Contribuição Científica

O desenvolvimento deste trabalho visa estudar maneiras de selecionar os genes mais relevantes para uma determinada característica, como também avaliar algoritmos de

agrupamento através de critérios multiobjetivo, auxiliando pesquisadores no processo da análise de bases contendo dados de expressões gênicas obtidas pela técnica de microarranjo.

1.3. Contribuição Técnica

É também disponibilizada uma ferramenta protótipo à comunidade científica especificamente aplicada à Mineração de dados e Aprendizagem de Máquina nos dados gerados pelo microarranjo. Com esta ferramenta é possível a realização de outras pesquisas e a comparação com resultados obtidos por pesquisas já realizadas, como em [Sunaga, 2006] e [Borges, 2006].

1.4. Estrutura do Trabalho

Este trabalho é dividido em 6 capítulos. No capítulo 2 são apresentados os conceitos básicos da seleção de atributos, sua importância na análise de bases de dados de expressões gênicas e os algoritmos usados neste trabalho. O capítulo 3 introduz a etapa do agrupamento e é mostrada sua influência na seleção de atributos, como também o funcionamento de alguns algoritmos e índices de validação de grupos pertencentes a diferentes abordagens. O capítulo 4 apresenta técnicas multiobjetivo e mostra como elas definem as melhores soluções para um problema de otimização multiobjetivo. No capítulo 5 é apresentada a proposta deste trabalho, onde serão descritas as etapas realizadas e as bases de dados utilizadas. Em seguida, no capítulo 6, são descritos os experimentos, juntamente com os resultados obtidos, e em seguida serão feitas as últimas considerações.

Por fim, este trabalho possui um apêndice contendo uma relação dos genes que mais foram selecionados pelos métodos de seleção de atributos que aqui foram abordados.

Capítulo 2

Seleção de Atributos

Muitas aplicações do mundo real, como análise de expressões gênicas obtidas pela técnica de microarranjo, apresentam grande quantidade de atributos [Dash et al., 2002]. O problema é que, em algumas dessas aplicações, muitos atributos não melhoram o resultado da análise e são tidos como redundantes ou irrelevantes [Dy, 2008]. Logo, para que os modelos sejam compreensíveis e que seus resultados sejam obtidos de forma mais simples, com altas taxas de acerto e em menos tempo, é necessário que essas aplicações passem por uma redução da dimensionalidade [Yu & Liu, 2004], [Liu & Yu, 2005], [Dy, 2008].

A técnica de redução de dimensionalidade abordada neste trabalho é a seleção de atributos, que pode ser aplicada tanto em problemas de aprendizagem supervisionada como de aprendizagem não-supervisionada e funciona como uma etapa de pré-processamento para outras etapas da mineração de dados, como a classificação e o agrupamento [Saeys, Inza & Larrañaga, 2007]. Ela é bastante usada em análises gênicas, em que as amostras contidas nas bases de dados ficam esparsas no espaço cartesiano, pois são formadas por muitas dimensões [Liu & Yu, 2005].

O objetivo da seleção de atributos é selecionar um subconjunto formado por atributos de uma base de dados a partir de um critério de avaliação, de forma que o modelo em uso se torne mais simples. Os atributos selecionados não devem sofrer alteração, extração ou construção e o subconjunto selecionado deve ser o mais representativo para a base de dados inteira [Kohavi & John, 1997], [Yu & Liu, 2004], [Saeys, Inza & Larrañaga, 2007].

Segundo [Yu & Liu, 2004], um conjunto de atributos representativo é aquele composto pelas características mais relevantes da base de dados. Diz-se que um atributo x_i é relevante se a classificação de uma classe Y é condicional ao valor de x_i dentro do espaço de

atributos s_i, ou seja, $x_i \in s_i$ (Equação 2.1). Uma classe é condicional a um atributo se a remoção deste atributo implica na perda da qualidade de classificação da classe.

$$p(Y_i = Y \mid X_i = x_i, S_i = s_i) \neq p(Y_i = Y, S_i = s_i) \quad (2.1)$$

Para [Kohavi & John, 1997], dentro da definição de relevância um atributo pode ser classificado como fortemente ou fracamente relevante. Ele é fortemente relevante caso sua remoção resulte em queda elevada de qualidade do conjunto de atributos selecionados, evitando assim que seja removido. Atributos são pouco relevantes se existir um subconjunto S_i de atributos, tal que o resultado obtido em S_i seja pior que em $S_i \cup \{x_i\}$.

Quando o atributo não se enquadra em nenhuma das definições anteriores, dizemos que este é irrelevante e desnecessário para o conjunto ótimo de atributos, já que não mostra dependência com nenhum dos demais atributos [Dy, 2008].

A depender de como seja escolhido o subconjunto de atributos relevantes, um atributo relevante pode ser substituído por um grupo de atributos que, em conjunto, trabalha melhor que o atributo sozinho. Por isso, diz-se que o nível de relevância de um atributo não é garantia de que este esteja no conjunto ótimo de atributos [Kohavi & John, 1997].

Além de relevantes, os atributos do conjunto ótimo não devem ser redundantes. Em [Yu & Liu, 2004], define-se atributos redundantes em um conjunto S quando um subconjunto de atributos $M_i \subset S$ é *Markov blanket* de um atributo S_i não pertencente a esse subconjunto. Um subconjunto é dito *Markov blanket* quando a probabilidade de sua ocorrência não é afetada pela eliminação de um determinado atributo. A ocorrência de um *Markov blanket* em um conjunto de classes Y dá-se a partir da Equação 2.2.

$$p(S - M_i - \{S_i\}, Y \mid S_i, M_i) = p(S - M_i - \{S_i\}, Y \mid M_i) \quad (2.2)$$

Sendo assim, o subconjunto ótimo de atributos em uma base de dados, como pode ser visto na Figura 2.1, é composto pelas características fortemente relevantes e pelas características pouco relevantes, porém não redundantes [Yu & Liu, 2004]. Este subconjunto não é único, já que podemos ter vários conjuntos com a mesma qualidade e com diferentes números de características [Kohavi & John, 1997].

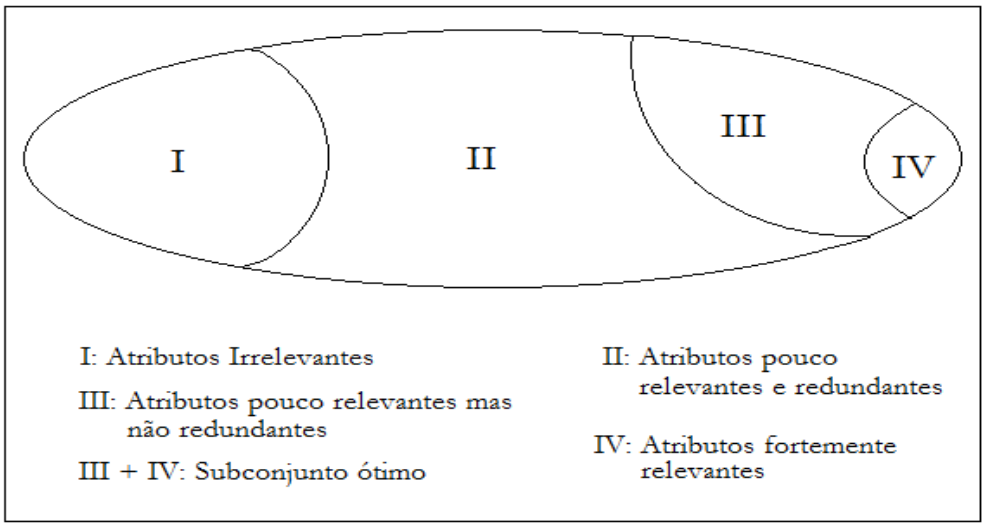

Figura 2.1: Tipos de atributos de uma base baseado em [Yu & Liu, 2004]

A seguir serão vistas as etapas do processo de seleção que, a depender de como estejam empregadas, definem o tipo do método. A classificação desses métodos, como será vista detalhadamente mais adiante, é comumente dividida entre métodos de filtro, que constroem modelos genéricos para qualquer algoritmo de mineração que venha ser utilizado mais adiante, ou métodos *wrapper*, que formam modelos para um algoritmo específico de mineração [Xing, Jordan & Karp, 2001], [Dy, 2008].

2.1. Etapas da Seleção de Atributos

Para conseguir gerar um subconjunto ótimo de atributos de uma base, que é o objetivo da seleção de atributos, é necessário o cumprimento de quatro etapas que auxiliarão na resolução deste problema. Elas estão representadas na Figura 2.2 e definem o modelo de seleção de atributos a depender de como o usuário as utiliza. A seguir, serão apresentadas as seguintes etapas: geração de subconjuntos, avaliação do subconjunto, critério de parada e validação do resultado.

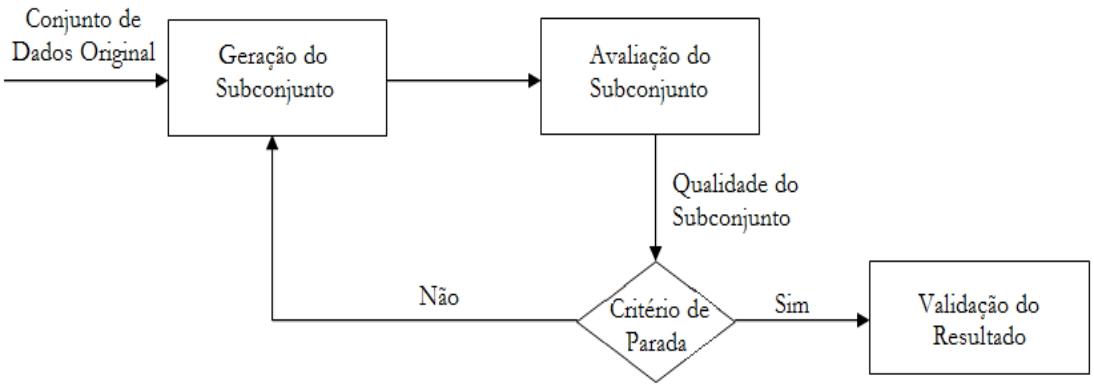

Figura 2.2: Etapas da seleção de atributos adaptado de [Dash & Liu, 1997]

2.1.1. Geração de Subconjuntos

Nesta etapa, o espaço de atributos é percorrido para determinar alguns subconjuntos de atributos que serão enviados para avaliação. Para isso, duas tarefas devem ser realizadas [Liu & Yu, 2005]:

- Definir o ponto de início. Um subconjunto ótimo pode ser iniciado como sendo um conjunto vazio que, ao decorrer da etapa, os atributos vão sendo adicionados, realizando assim uma seleção *Forward*. Pode-se fazer também uma seleção *Backward*, onde o subconjunto é iniciado com todas as características e aquelas que não forem consideradas relevantes são removidas. Apesar de ser mais fácil, a seleção *Backward* é extremamente mais cara computacionalmente e mais lenta [Kohavi & John, 1997].

- Decidir uma estratégia de busca. Para garantir que o resultado retornado seja um subconjunto ótimo, uma busca completa deve ser feita [Dy, 2008]. Porém, esta busca pode acabar sendo inviável, pois quando o espaço de características cresce, a quantidade de combinações de atributos cresce exponencialmente. Para isso, funções heurísticas podem ser empregadas para redução do espaço de busca sem diminuir as chances de obter êxito no resultado [Saeys, Inza & Larrañaga, 2007]. Uma busca de implementação mais simples e que produz resultados mais rapidamente é a busca sequencial, mas ela não garante o retorno do resultado ótimo [Dy, 2008]. Pode-se realizar também uma busca aleatória iniciando com um subconjunto sorteado de forma aleatória e seguir de duas formas: combinando busca sequencial com busca aleatória, ou gerar o próximo subconjunto de maneira totalmente aleatória [Liu & Yu, 2005].

2.1.2. Avaliação de Subconjuntos

Após a geração dos subconjuntos de atributos é necessário que estes sejam avaliados para certificar que um subconjunto válido, podendo ser ótimo, foi encontrado [Kohavi & John, 1997]. Os critérios de avaliação estão classificados, segundo [Liu & Yu, 2005], de acordo com a dependência com os algoritmos de mineração.

- Critério independente: recebe esse nome por não envolver um algoritmo de mineração durante a avaliação. É tipicamente usado pelo método de filtro, que será visto mais adiante, pois só explora as características intrínsecas. Entre as medidas utilizadas no critério independente encontra-se a medida de distância, em que são selecionados os atributos que melhor separam as classes [Liu & Yu, 2005]. Em [Dash et al., 2002] é utilizada uma medida baseada na entropia (desordem) de um sistema.

- Critério dependente: diferentemente do critério independente, um algoritmo de mineração pré-determinado é combinado com o critério de avaliação. Muito utilizado em métodos *Wrappers*, este critério é mais caro computacionalmente e específico para o algoritmo utilizado [Liu & Yu, 2005].

2.1.3. Critério de Parada

A etapa de geração de subconjuntos é realizada de forma iterativa até que algum critério de parada estabelecido seja alcançado [Liu & Yu, 2005]. Um critério de parada é atingido, por exemplo, quando:

- O final da busca é atingido. No caso da busca completa, ao explorar a última combinação de atributos, a seleção retorna os melhores subconjuntos;
- Limite de iteração ou quantidade de atributos mínima é alcançado;
- Mudanças na configuração do algoritmo de mineração não geram subconjuntos melhores;
- Um subconjunto considerado ótimo pelo desempenho do algoritmo de mineração é encontrado.

2.1.4. Validação do Resultado

Assim como em qualquer etapa da mineração de dados, o resultado obtido na seleção de atributos deve ser validado. A forma mais fácil de julgar a qualidade do subconjunto

gerado é tendo conhecimento prévio suficiente sobre a base de dados e sobre o problema [Liu & Yu, 2005].

O único conhecimento *a priori* na análise de uma base de expressões gênicas é saber a qual característica as amostras estavam submetidas no momento de serem obtidas pelo microarranjo. Desta forma, este trabalho mede a qualidade dos genes selecionados através do agrupamento, verificando se amostras de pessoas portadoras da mesma característica se comportam de forma similar.

2.2. Método de Seleção Tipo Filtro

A partir de uma base de dados D, seus atributos podem ser explorados e avaliados sem que haja interferência de um algoritmo de mineração, ou seja, as características são filtradas independentemente de um modelo [Kohavi & John, 1997].

O subconjunto de atributos S_{best}, representado no Algoritmo 2.1, é a melhor resposta corrente e é a solução retornada ao atingir o critério de parada δ [Liu & Yu, 2005]. Ele pode ser iniciado como sendo S_0 um conjunto vazio, ou com todos os atributos ou ainda sorteando aleatoriamente alguns atributos [Liu & Yu, 2005].

Ao final de toda iteração, um subconjunto é gerado com os atributos mais relevantes até o momento. Esse subconjunto é comparado com S_{best} por meio de uma medida independente M de avaliação e, caso este seja melhor, passará a ser o novo S_{best} [Saeys, Inza & Larrañaga, 2007].

Algoritmo 2.1 Método Filtro

```
Inicializar S_best = S_0
Avaliar S_best com M
Enquanto critério_parada não for alcançado
    Geração do subconjunto de atributos S_atual
    Avaliar S_atual com M
    Se S_atual melhor que S_best
        S_best = S_atual
Fim enquanto
Retorna S_best
```

A fim de abordar diferentes buscas pelo espaço de atributos, este trabalho fará uso dos seguintes métodos filtro: C-*Focus*, *Relief*-F e CFS.

2.2.1. C-Focus

Um algoritmo de seleção bastante usado é o *Focus* (Algoritmo 2.2), que faz uso de um sistema de busca completa pelos N atributos da base, em que todas as combinações de subconjuntos de atributos são examinadas, eliminando inconsistências ou redundâncias apenas no subconjunto em análise e retorna o menor subconjunto ótimo, que é chamado de *Min-Features* [Arauzo, Benitez & Castro, 2003]. O problema desse algoritmo para este trabalho é que o *Focus* não lida com atributos numéricos, que são usadas na análise de expressões gênicas e, por isso, este trabalho faz uso do algoritmo C-*Focus*, que é uma modificação do método *Focus* para trabalhar com atributos numéricos.

Algoritmo 2.2 Focus

```
Inicializar Min_Features = S0
Para i = 1 até N
      Verificar inconsistências em cada subconjunto L
      Se não houver inconsistências em L
            Min_Feature = L
Fim Para
Retorna Min_Features
```

2.2.2. Relief-F

O método de seleção de atributos *Relief* (Algoritmo 2.3) tenta estimar uma quantidade K de atributos especificada pelo usuário. Quanto maior o poder de um atributo em distinguir instâncias de classes diferentes, maior é a probabilidade dele ser selecionado. A limitação do *Relief* é de suportar, no máximo, apenas duas classes [Kononenko & Sikonja, 2008].

Ele executa uma busca aleatória e, para cada instância Ins sorteada devem ser achadas as instâncias *hit*, que são a mais próximas e pertencentes a mesma classe, e as instâncias *miss*, que são as mais próximas e pertencentes a classes diferentes. Com essas instâncias é calculado o vetor de relevância $W[A]$ (Equação 2.3) para cada atributo A que a compõe, em que m é a quantidade de instâncias na base.

A equação de $diff$ (Equação 2.4) retorna a diferença do valor entre as instâncias ins_1e ins_2 para o atributo A. O uso do maior e menor valor de A, representados respectivamente por $max(A)$ e $min(A)$, serve como forma de normalização para que o resultado de $diff$ fique no intervalo [0,1].

$$W[A] = W[A] - \frac{diff(A,Ins,Hit)}{m} + \frac{diff(A,Ins,Miss)}{m} \quad (2.3)$$

$$diff(A, Ins_1, Ins_2) = \frac{|Ins_{1,att(A)} - Ins_{2,att(A)}|}{\max(A) - \min(A)} \quad (2.4)$$

O resultado retornado é o conjunto dos k atributos com os maiores valores de relevância [Kononenko & Sikonja, 2008].

O método *Relief*-F é uma extensão do método *Relief* que abrange mais de duas classes e lida também com dados incompletos. A única diferença é no cálculo do vetor de relevância W, representada pela Equação 2.5, pois são procuradas as instâncias mais próximas para cada uma das C classes existentes [Kononenko & Sikonja, 2008].

$$W[A] = W[A] - \frac{diff(A,Ins,Hit)}{m} + \sum_{i=1}^{C} \frac{diff(A,Ins,Miss_i)}{m} \quad (2.5)$$

Algoritmo 2.3 Relief

```
Inicializar W
Para i = 0 até (m - 1)
     Selecionar uma instância Ins aleatoriamente
     Procurar pelas instâncias hit e miss de Ins
     Para j = 0 até (A - 1)
          Atualizar W[A]
Fim Para
W_aux = ordena(W[A]) //ordena W[A] e escolhe os K maiores
Retorna W_aux
```

2.2.3. CFS

O método CFS, *Correlation-based Feature Selection*, gera subconjuntos de atributos baseados na correlação deles com a classe a que pertencem. Dois atributos se correlacionam se, trabalhando juntos, conseguem predizer uma classe mais facilmente que se estiverem trabalhando individualmente [Hall, 2000].

Primeiramente é calculada uma matriz de correlação classe-atributo e atributo-atributo. A qualidade de um subconjunto S, formado por K atributos, é medida pelo seu mérito representado pela Equação 2.6, em que $\overline{r_{cf}}$ é a média das correlações classe-atributo do subconjunto e $\overline{r_{ff}}$ é a média das correlações atributo-atributo.

$$mérito_S = \frac{K \times \overline{r_{cf}}}{\sqrt{K + K \times (K-1) \times \overline{r_{ff}}}} \tag{2.6}$$

O CFS inicia o subconjunto S sem nenhum atributo, como é visto no Algoritmo 2.4, e através de uma busca sequencial, vai adicionando atributos e expandindo o subconjunto que obtiver o melhor mérito.

Este método é facilmente usado por atributos discretos e contínuos e realiza uma classificação de subconjuntos de atributos, diferentemente do *Relief*-F, que classifica os atributos individualmente. Os subconjuntos selecionados tendem a possuir baixa redundância e alto poder de caracterizar uma classe [Hall, 2000].

Algoritmo 2.4 CFS

```
Inicializar S
merito_final = merito(S)
merito_maior = true
Enquanto (merito_maior)
     merito_atual = 0;
     Para K = 0 até (m - 1)
          Se atributo[K] não está em S
               Se merito(S + atributo[K]) > merito_atual
                    atributo_add = atributo[K]
```

```
                    merito_atual = merito(S + atributo[K])
                Fim Se
        Fim Para
        Se merito_atual > merito_final
            merito_final = merito_atual
            Adiciona atributo_add a S
            merito_maior = true
        Senão
            merito_maior = false;
Fim Enquanto
Retorna S
```

2.3. Método *Wrapper*

Uma seleção de atributos, diferentemente do método de filtro, pode ser específica para um algoritmo de mineração se este for envolvido durante a formação do subconjunto ótimo [Liu & Yu, 2005]. O método *wrapper*, apresentado no Algoritmo 2.5, se encarrega de fazer isso atingindo melhores performances comparados aos métodos de filtro e levando em consideração a existência de dependências entre as características [Saeys, Inza & Larrañaga, 2007]. Por outro lado, não funciona bem com bases de dados que contêm muitos atributos, pois acaba sendo mais custoso computacionalmente pelo fato de que o algoritmo de mineração, também chamado de algoritmo de indução, é executado várias vezes nas validações dos subconjuntos candidatos, a fim de induzir o método de seleção a escolher o subconjunto que melhor trabalha com o método de mineração [Xing, Jordan & Karp, 2001].

Já que uma seleção de atributos pode ser utilizada tanto para aprendizagem supervisionada quanto para a não-supervisionada, o algoritmo de indução pode pertencer à etapa de classificação ou à de agrupamento e deve ser aquele que será usado nas próximas etapas [Saeys, Inza & Larrañaga, 2007] [Dy, 2008].

Algoritmo 2.5 Método Wrapper

```
Inicializar S_best = S_0
Avaliar S_0 com um algoritmo de mineração A
Enquanto critério_parada não for alcançado
```

Geração do subconjunto de atributos S_{atual}

Avaliar S_{atual} com A

Se S_{atual} **melhor que** S_{best}

$S_{best} = S_{atual}$

Fim enquanto

Retorna S_{best}

2.4. Considerações Finais

Este capítulo apresentou a técnica de seleção de atributos, cuja finalidade é reduzir a dimensionalidade de uma base de dados através da seleção de um subconjunto formado pelos atributos mais relevantes dessa base.

Os métodos de seleção de atributos são comumente divididos em métodos filtro e métodos *wrappers*. Esse último, por utilizar repetidas vezes um algoritmo de indução, acaba sendo custoso e desvantajoso ao se tratar de bases de dados que contêm muitos atributos.

A fim de abordar diferentes buscas pelo espaço de atributos, este trabalho fará uso dos seguintes métodos de filtro: C-*Focus*, *Relief*-F e CFS. A qualidade dos métodos de seleção de atributos, já que podem ser aplicadas em problemas não-supervisionado, será medida na etapa de agrupamento. O objetivo do agrupamento neste trabalho, que será visto no capítulo seguinte, é verificar se pessoas portadoras da mesma característica se comportam de forma similar. Dessa forma, o melhor agrupamento identificará o melhor método de seleção de atributos e será possível descobrir quais genes são mais relevantes para uma determinada característica.

Capítulo 3

Agrupamento

A técnica de agrupamento tende a organizar um conjunto de objetos em grupos, de forma que aqueles de comportamentos mais parecidos fiquem em um mesmo grupo, com a finalidade de revelar a estrutura natural das bases de dados [Xu & Wunsch, 2005], [Dy, 2008]. No caso deste trabalho, o agrupamento de instâncias de expressões gênicas ajudará a identificar aquelas que compartilham as mesmas características.

Um bom agrupamento ocorre quando os atributos que formam as bases de dados são relevantes e suas instâncias definem bem os grupos aos quais pertencem. Desse ponto de partida, um conjunto de genes selecionados como relevantes podem resultar em um bom agrupamento. Por essas razões, o agrupamento pode servir como avaliador dos métodos de seleção de atributos e revelar os genes que estão envolvidos com a aparição de uma determinada doença, tornando-se um método comum na análise genética [Quackenbush, 2001].

Diferentemente da classificação tradicional, que faz uso de grupos existentes e definidos na fase de treinamento para classificar de forma supervisionada novos objetos, o agrupamento faz a criação de grupos a partir de uma coleção de objetos sem rótulos, sendo assim chamado também de classificação não supervisionada [Jain, Murty & Flynn, 1999].

As etapas básicas que devem compor o processo de agrupamento podem ser divididas em [Jain & Dubes, 1988]:

a) Primeiramente deve-se escolher a representação dos objetos e a quantidade de exemplos disponíveis, assim como os tipos e tamanhos de seus atributos. Opcionalmente pode-se realizar uma seleção manual de atributos, principalmente se houver uma grande quantidade de características, que pode diminuir o trabalho e simplificar o processo;

b) Selecionar um algoritmo de agrupamento para aplicar aos objetos relacionados. Essa tarefa pode ser auxiliada pela escolha da função critério que deseja realizar para verificar a proximidade entre padrões. A escolha do algoritmo é de extrema importância, já que não há um único algoritmo que possa servir para qualquer situação, e talvez nunca haja, pois existem particularidades como: tipo de agrupamento a ser feito, tipo dos grupos resultantes e medida de proximidade a ser aplicada [Quackenbush, 2001];

c) Se um algoritmo gera grupos diferentes de outro, então é necessário que haja uma forma de avaliar esses grupos para sabermos qual algoritmo se comporta melhor, reforçando a importância da escolha deste. A forma de avaliação pode ser do tipo externa, interna ou relativa, e serão vistas com maior detalhe mais adiante;

d) A concretização de um bom agrupamento se dá no momento em que os grupos resultantes têm algum significado. Nessa etapa é importante a interpretação dos resultados por alguém que tenha conhecimento sobre os dados;

A divisão dos tipos de agrupamento mais utilizada é entre hierárquico e particional (Figura 3.1). A seguir, esses tipos serão apresentados, assim como os métodos usados neste trabalho. Em seguida, serão vistos os chamados índices de validação de grupos, que servem para qualificar um agrupamento.

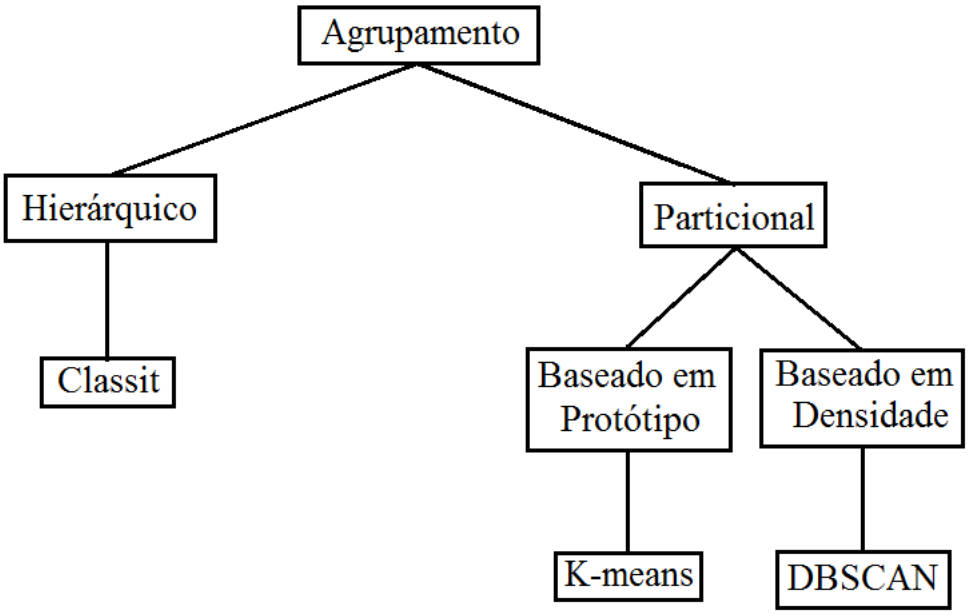

Figura 3.1: Classificação dos tipos de agrupamento baseado em [Jain, Murty & Flynn, 1999]

3.1. Agrupamento Hierárquico

Como o próprio nome diz, esse método organiza os grupos gerados de forma hierárquica, permitindo que um único objeto pertença a vários grupos que compartilham semelhanças, mas que também têm algumas características diferentes. Sua representação pode ser feita através de um dendograma (Figura 3.2a), mostrando a disposição final para cada objeto em forma de árvore, ou de um diagrama aninhado (Figura 3.2b) [Tan, Steinbach & Kumar, 2006].

Na Figura 3.2 é possível visualizar que um grupo é composto pelos objetos p2 e p3. Adicionando o objeto p4, um grupo diferente é formado pelos três objetos. Um terceiro grupo é formado ao se incluir o objeto p1.

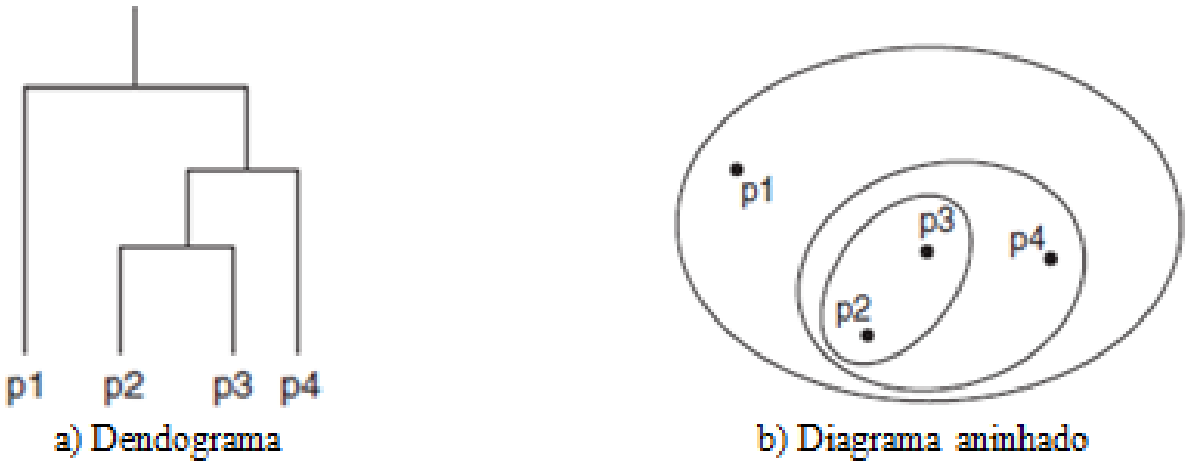

Figura 3.2: Representações hierárquicas [Tan, Steinbach & Kumar, 2006]

Para uma base dados não há apenas um único modo de agrupar seus objetos hierarquicamente. A construção da estrutura final depende de como eles serão tratados inicialmente, de forma aglomerativa ou divisiva. A forma aglomerativa, que é a forma mais usada, inicia com os objetos de forma independente e, no decorrer do processo, os objetos vão sendo unidos e formando grupos maiores. Na forma divisiva, os objetos pertencem a um único grupo e estes vão se diferenciando uns dos outros e formando subgrupos [Tan, Steinbach & Kumar, 2006].

O método de agrupamento hierárquico abordado neste trabalho será o Classit, que realiza um agrupamento conceitual nas bases de dados e lida com atributos numéricos, que é o caso dos atributos da bases de expressões gênicas.

3.1.1. Classit

O agrupamento conceitual enquadra uma instância a depender do comportamento individual de cada atributo, de forma que cada conceito é composto pelas instâncias que compartilham comportamentos similares de um conjunto de atributos. Divergências entre duas instâncias podem acarretar a criação de um novo conceito mais especializado e o conjunto dos conceitos situados no mesmo nível da estrutura hierárquica formam uma partição.

Cobweb é um algoritmo popular de agrupamento conceitual. A classificação das X instâncias do banco na hierarquia da árvore é construída por intermédio de quatro operadores [Devaney & Ram, 1997]:

- Que incorpora uma instância a um conceito existente;
- Que cria um novo conceito e abriga a instância;
- Que une dois grupos para generalizar os conceitos;
- Que divide grupos para especializar os conceitos.

A melhor operação em um determinado momento é aquela que obtiver o melhor valor da função de avaliação. O incremento de uma instância tem fim quando esta alcança uma folha ou quando é criado um novo conceito na estrutura hierárquica.

O problema do Cobweb é que a função de avaliação não lida com valores numéricos, que é o caso das características de expressões gênicas. Para isso, pode-se usar o método Classit (Algoritmo 3.1), que é baseado no Cobweb e classifica instâncias formadas por atributos numéricos em conceitos hierárquico.

A função de avaliação (FA) é dada pela Equação 3.1. Nela, n é o número de classes em uma partição, σ_{ik} é o desvio padrão do atributo i na classe k e σ_{pi} é o desvio padrão do atributo i no nó pai. O termo C_k é o conceito da partição k.

$$FA = \frac{\sum_k^n P(C_k) \times \frac{1}{\sigma_{ik}} - \frac{1}{\sigma_{pi}}}{n} \tag{3.1}$$

Algoritmo 3.1 Classit

```
Inicializar Arvore = vazio (Passo 1)
For m = 1 até X (Passo 2)
     Se raiz for vazio (Passo 3)
          Arvore.raiz = ins[m]
     Senão (Passo 4)
          noAtual = raiz
          Se noAtual = folha (Passo 5)
               Criar conceito com ins[m]
               Arvore.noAtual.filho = ins[m]
          Senão (Passo 6)
               A = Melhor FA dos filhos de noAtual
               B = Segundo melhor FA dos filhos de noAtual
               C = FA se ins[m] for filho de noAtual
               D = FA para união de A e B
               E = FA para divisão de A
               Se A é melhor FA calculado
                    noAtual = noAtual.filho_maior_FA
                    Repetir Passo 5
               Se C é melhor FA calculado
                    Adicionar ins[m] como filho de noAtual
                    Arvore.noAtual.filho = ins[m]
               Se D é melhor FA calculado
                    novoGrupo = união de A e B
                    noAtual = novoGrupo
                    Repetir Passo 5 //Agora com um novo grupo
               Se E é melhor FA calculado
                    Dividir A
                    Repetir Passo 5 //Agora com grupos novos
          Fim Senão
Fim For
Retorna Arvore
```

3.2. Agrupamento Particional

O objetivo do agrupamento particional é gerar uma partição de grupos sem relações hierárquicas entre eles, ou seja, um objeto pertence apenas a um grupo e não a uma estrutura de agrupamento (como o dendograma, no caso do modo hierárquico) [Jain, Murty & Flynn, 1999], [Xu & Wunsch, 2005]. Por esse fato, a realização de um agrupamento particional acaba sendo mais adequado para bases de dados grandes, como é o caso das bases de expressões gênicas.

Por outro lado, o agrupamento hierárquico pode obter melhores resultados que o agrupamento particional, já que é específico para um determinado modelo.

Como dito na segunda etapa básica do agrupamento, métodos de agrupamentos diferentes geram grupos de tipos diferentes. Segundo [Tan, Steinbach & Kumar, 2006], entre os tipos de grupos que podem ser formados estão:

- Bem-separados, em que qualquer objeto de um grupo é mais similar com um objeto do próprio grupo do que com um situado em outro grupo (Figura 3.3a);
- Baseados em densidade, no qual um grupo é definido como uma região densa de objetos. Na Figura 3.3b, o círculo branco e o arco mais escuro são grupos, por terem grande concentração de objetos, que estão rodeados por uma região de baixa densidade;
- Baseados em protótipos, quando um objeto pertence ao grupo cujo protótipo está mais próximo (a definição de protótipo pode ser representada por um centróide).

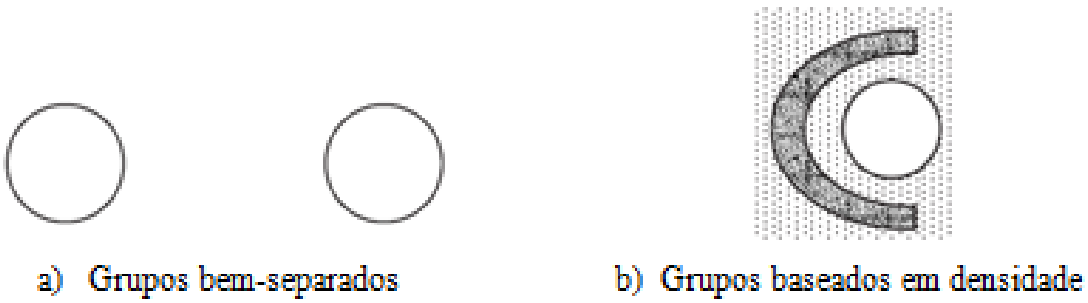

Figura 3.3: Tipos de grupos [Tan, Steinbach & Kumar, 2006].

Os métodos de agrupamento particional usados neste trabalho, por serem simples e terem obtidos bons resultados, são o K-*means* e o ISODATA. Eles geram grupos baseados em protótipos mas possuem diferenças que torna o ISODATA uma evolução do K-*means*.

3.2.1. K-means

O *K-means* é um método particional bastante conhecido e simples que é baseado em protótipo, ou seja, os objetos da base de dados são associados ao centróide, ou protótipo, mais próximo. O objetivo desse método é formar K grupos, sendo este valor especificado pelo usuário, assumindo que eles tenham formato hiper-esférico e definindo os centróides como seus respectivos pontos centrais [Tan, Steinbach & Kumar, 2006].

A escolha dos centróides iniciais é feita de forma aleatória e a associação dos objetos é feita pela soma do erro quadrático, ou SSE (*Sum of Square-Error*). Ela, representada pela Equação 3.2, calcula a soma das distâncias euclidianas $dist(x_i, x)$ das d dimensões entre o objeto x_i e um referencial do grupo, que pode ser um centróide x, apresentado na Equação 3.3. Dessa forma, o objeto pertencerá ao grupo cujo centróide está mais próximo.

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x_i, x)^2 \tag{3.2}$$

$$dist(x_i, x)^2 = \sum_{l=1}^{d} |x_{il} - x_l|^2 \tag{3.3}$$

Ao final de cada iteração será realizada a atualização dos pontos centrais dos grupos através da Equação 3.4, em que o centróide C do grupo i é formado pela média dos atributos x que formam os m objetos c. O passo da associação é repetido até que não haja grandes modificações na atualização dos centróides ou a quantidade de iterações é atingido [Xu & Wunsch, 2005].

$$C_i = \frac{\sum_{x \in c_i} x}{m_i} \tag{3.4}$$

O problema do K-*means*, representado pelo Algoritmo 3.2, é não identificar a quantidade de grupos naturais e formar o valor especificado pelo usuário [Xu & Wunsch, 2005]. O algoritmo ISODATA, que será visto a seguir, é capaz de realizar um auto-ajuste na quantidade de grupos, tornando-se uma evolução do K-*means*.

Algoritmo 3.2 K-means

```
Inicializar K centróides selecionando pontos aleatórios
parada = false
Enquanto (parada = false)
     For m = 1 até X (quantidade de instâncias)
          Agrupamento = Associar inst[m] ao centróide mais próximo
     Fim For
     Recomputar centróides
     Se (centróides próximos || última iteração)
          parada = true
Fim Enquanto
Retorna Agrupamento
```

3.2.2. ISODATA

O *Iterative Self-Organizing Data Analysis Technique* (Algoritmo 3.3), mais conhecido como algoritmo de agrupamento ISODATA, é uma implementação do K-*means* que faz a identificação de grupos compactos através de um auto-ajuste da quantidade de grupos em um agrupamento [Theodoridis & Koutroumbas, 2003].

Essa adaptação do K-*means* tem a capacidade de fazer divisão ou união de grupos formados a depender da quantidade de objetos que um grupo pode suportar [Arai & Bu, 2007]. Se um grupo tiver uma quantidade de objetos abaixo do limiar mínimo, este desaparecerá e seus objetos serão reagrupados nos grupos restantes mais próximos, mas se houver um número de objetos superior ao limiar máximo, este grupo será dividido [Jain & Dubes, 1988]. Um terceiro caso acontece quando dois ou mais centróides estão muito próximos um do outro e são unidos [Arai & Bu, 2007].

O critério de parada, também como no K-*means*, é a baixa taxa de modificação do agrupamento ao final de uma iteração, assim como o número de iterações limite, estabelecida pelo usuário [Arai & Bu, 2007].

Algoritmo 3.3 ISODATA

```
Inicializar K centróides selecionando pontos aleatórios
parada = false
Enquanto (parada = false)
     For m = 1 até X (quantidade de instâncias)
          Agrupamento = Associar inst[m] ao centróide mais
próximo
     Fim For
     For i = 1 até K
          Se grupo[i].tamanho < mínimo_instâncias
               Redistribuir instancias de grupo[i]
               Remover grupo[i]
          Fim Se
          Se grupo[i].tamanho > máximo_instâncias
               Dividir grupo[i]
          Fim Se
          Se grupo[i].centroide perto centróide de outro grupo
               Unir grupos
          Fim Se
     Fim For
     Recomputar centróides
     Se (centróides próximos || última iteração)
          parada = true
Fim Enquanto
Retorna Agrupamento
```

3.3. Validação de Grupos

Pelo fato de haver métodos de agrupamentos que consideram diferentes conceitos de grupos e acabam obtendo resultados diferentes, é importante utilizar uma forma de validar esses agrupamentos, a fim de certificar se o resultado é bom ou ruim.

Os critérios de validação de grupos objetivam identificar a quantidade correta de grupos em uma base de dados, saber se os agrupamentos gerados têm significado e revelar qual método de agrupamento se comportou melhor.

Há três tipos de índices de validação de grupos, que serão vistos a seguir: relativos, internos e externos.

3.3.1. Índice Relativo

Segundo [Faceli, Carvalho & Souto, 2005], a utilização mais comum dos índices relativos é para determinar o número adequado de grupos. Para isso, executa-se o algoritmo de agrupamento para a quantidade K que se deseja formar, o número K_{min} mínimo de grupos e o número K_{max} máximo de grupos que podem ser formados.

- Índice C: A obtenção do valor C deste índice, Equação 3.5, é feita através da razão envolvendo a soma das distâncias S_U de p pares de amostras pertencentes ao mesmo grupo U, p menores distâncias e p maiores distâncias entre todas as amostras da base de dados, representadas por S_{min} e S_{max}, respectivamente [Bolshakova & Azuaje, 2006].

$$C(U) = \frac{S_U - S_{min}}{S_{max} - S_{min}} \tag{3.5}$$

Para este índice, baixos valores correspondem a bons grupos, pois representam baixas distâncias internamente, e a quantidade K de grupos cuja soma de todos os índices C seja mínima, será o número ótimo de grupos para o agrupamento realizado [Bolshakova & Azuaje, 2006].

- Índice Dunn: É a razão da separação interna nos grupos e entre os grupos, tornando-se sensível a ruídos (Equação 3.6). Grandes valores para este índice correspondem a bons grupos e o valor de K que o maximiza é o número ótimo de grupos [Faceli, Carvalho & Souto, 2005], [Maulik & Bandyopadhyay, 2002].

$$D(K) = \min_{i=1,\dots,K}\left\{\min_{j=i+1,\dots,K}\left\{\frac{\delta(C_i,C_j)}{\max_{K=1,\dots,K} diametro(C_K)}\right\}\right\} \tag{3.6}$$

$$\delta(C_i, C_j) = \min_{x\in C_i, y\in C_j} d(x,y) \tag{3.7}$$

$$diametro(C_K) = \max_{x,y\,\in\, C_i} d(x,y) \tag{3.8}$$

A expressão $\delta(C_i, C_j)$, Equação 3.7, representa uma função de dissimilaridade entre grupos, ou seja, a menor distância entre pontos, sendo que um ponto pertence ao grupo C_i e o outro ponto pertence ao grupo C_j, e o $diametro(C_K)$, Equação 3.8, retorna o diâmetro do grupo C_K, ou seja, a maior distância entre pontos do mesmo grupo.

- Índice Davies-Bouldin: Esse índice, bastante aplicável em agrupamentos hiper-esféricos, é a razão da soma da dispersão interna nos grupos pela separação entre grupos, de acordo com as equações 3.9, 3.10 e 3.11. Sua tarefa é minimizar o valor do índice para atingir um agrupamento ideal, representado pelo valor 0 [Faceli, Carvalho & Souto, 2005], [Maulik & Bandyopadhyay, 2002].

$$DB(K) = \frac{1}{K}\sum_{x=1}^{K} R_x \tag{3.9}$$

$$R_x = \max_{j \neq x}\{R_{j,x}\} \tag{3.10}$$

$$R_{j,x} = \frac{e_{C_j} + e_{C_x}}{d(C_j, C_x)} \tag{3.11}$$

$R_{j,x}$, na Equação 3.9, é a função de similaridade entre dois grupos medida pela soma dos erros médios desses grupos dividida pela distância dos centros desses grupos e R_x. A Equação 3.8 calcula a dispersão do grupo C_x .

- Índice Silhueta: Neste índice podem ser usadas medidas de similaridade, sendo a distância Euclidiana a mais comum. A silhueta $s(x_i)$ para um ponto x_i é dada pela relação 3.12 [Faceli, Carvalho & Souto, 2005]:

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & a(x_i) < b(x_i) \\ 0, & a(x_i) = b(x_i) \\ a(x_i)/b(x_i) - 1, & a(x_i) > b(x_i) \end{cases} \tag{3.12}$$

O valor de uma silhueta está no intervalo [-1,1], sendo que o valor máximo representa um bom agrupamento para o ponto x_i e o valor mínimo, o contrário [Bolshakova & Azuaje, 2003]. O índice Silhueta tem melhor funcionamento em agrupamentos de formato esféricos, como o K-*means* e o ISODATA.

Para casos de similaridade, $b(x_i)$ é a similaridade média do ponto x_i para os outros pontos pertencentes ao mesmo grupo, a(x_i) é a maior similaridade média de x_i em relação a todos os demais grupos, em que $d(x_i, C_j)$ é a similaridade média do ponto x_i para todos os pontos do grupo C_j (Equação 3.13).

$$a(x_i) = \max_{C_i \neq C_j} d(x_i, C_j) \tag{3.13}$$

A soma das silhuetas de todos os pontos de um grupo representa o valor da silhueta para este grupo e a média $\bar{s}(K)$ das silhuetas de todos os K grupos representa a qualidade do agrupamento realizado, sendo que o K que obtiver o maior valor de $\bar{s}(K)$, chamado de coficiente silhueta, ou SC (*Silhouette Coeficient*) será a quantidade ótima de grupos para a base de dados [Azuaje & Bolshakova, 2002].

Segundo [Faceli, Carvalho & Souto, 2005], a interpretação para o valor do SC é feita da sguinte maneira:

- Se $SC \leq 0.25$, o agrupamento realizado é não substancial;
- Se $0.26 \leq SC \leq 0.5$, o agrupamento realizado é fraco;
- Se $0.51 \leq SC \leq 0.7$, o agrupamento realizado é razoável;
- Se $0.71 \leq SC \leq 1$, o agrupamento realizado é bom;

3.3.2. Índice Interno

Um critério interno, segundo [Faceli, Carvalho & Souto, 2005], mede a qualidade de uma partição gerada através dos dados empregados, usando sua matriz de similaridade. Além dos critérios específicos desse tipo de validação, como o *isolation*, que será visto a seguir, pode-se fazer uso dos exemplos de índices relativos citados como índices internos.

- **Isolation**: Muitas medidas de validação de grupos para agrupamento particional são baseadas no coeficiente chamado *isolation*. Esta medida é calcula pela proximidade dos centros de dois grupos [Tan, Steinbach & Kumar, 2006]. A medida de proximidade usada neste trabalho e que já foi citada é o SSE.

O critério de separação, como também pode ser chamado esse índice, tem como objetivo minimizar a SSE. Nele é calculada a separação entre um grupo, representado pelo seu centróide, e o centróide global c (Equação 3.14) [Tan, Steinbach & Kumar, 2006].

$$isolation(C_i) = SSE(c_i, c) \quad (3.14)$$

3.3.3. Índice Externo

Um critério supervisionado verifica, a partir do resultado obtido, se o agrupamento realizado é apropriado para os dados da base. Essa verificação é feita através da comparação de uma partição P_e gerada pelo algoritmo de agrupamento com outra partição independente, P_r, gerada com base em conhecimento a priori sobre a estrutura real dos dados.

Segundo [Faceli, Carvalho & Souto, 2005], a comparação de um par de pontos pertencentem a P_e e P_r pode ser feita da seguinte forma:

- Se os pontos pertencem ao mesmo grupo em P_e e P_r, é classificado como SS;

- Se os pontos pertencem ao mesmo grupo em P_e, mas pertencem a grupos diferentes em P_r, é classificado como SD;

- Se os pontos pertencem a grupos diferentes em P_e, mas pertencem ao mesmo grupo em P_r, é classificado como DS;

- Se os pontos pertencem a grupos diferentes tanto em P_e quanto em P_r, é classificado como DD;

- **Índice Jaccard (J)**: É a razão da quantidade de pares SS pela soma da quantidade de pares que pertençam ao mesmo grupo em pelo menos um agrupamento, ou seja (Equação 3.15):

$$J = \frac{a_1}{a_1 + a_2 + a_3} \quad (3.15)$$

Os valores a_1, a_2, a_3, a_4 representam, respectivamente, a quantidade de pares SS, SD, DS, DD [Faceli, Carvalho & Souto, 2005]. O valor máximo que se pode ter é 1, quando todos os pares são SS, alcançando um ótimo agrupamento.

3.4. Considerações Finais

Neste capítulo foi apresentada a etapa de agrupamento, que organiza de forma não supervisionada os objetos com comportamentos mais parecidos em um mesmo grupo. Foi mostrado que essa etapa pode servir como avaliador dos métodos de seleção de atributos e revelar os genes que estão envolvidos com a aparição de uma determinada doença.

A divisão mais usada é entre agrupamento hierárquico e particional. O primeiro, por construir estruturas aninhadas contendo os objetos, acaba se tornando inadequado para agrupar bases de dados com uma grande quantidade de atributos, mas ao utilizar bases reduzidas pode gerar bons agrupamentos. Já o agrupamento particional, que não constrói estruturas além das próprias partições, é mais adequado para bases de expressões gênicas.

O Classit é o método hierárquico conceitual utilizado neste trabalho. Já os métodos de agrupamento particional usados são o K-*means*, bastante popular e que gera grupos baseados em protótipo, e o ISODATA, que realiza um auto-ajuste da quantidade de grupos e é considerado evolução do K-*means*.

Com a formação de diferentes tipos de grupos, é importante uma forma de avaliar os resultados para certificar se o mesmo é bom ou ruim, além de identificar a quantidade de grupos em uma base de dados. Há três tipos de índices de validação de grupos abordados neste trabalho: relativos, internos e externos.

O uso de vários índices de validação serve para que não haja uma avaliação tendenciosa do agrupamento. Por essa razão torna-se necessária uma otimização dos vários índices de forma simultânea e isso é possível com o uso das técnicas de otimização multiobjetivo, que serão vistas no capítulo a seguir.

Capítulo 4

Técnicas Multiobjetivo

Para que um agrupamento tenha validade é necessário que os grupos resultantes tenham algum valor significativo. Para isso, critérios de validação de grupos tornam-se indispensáveis para o processo de agrupamento. Porém, uma escolha ruim do critério pode fazer com que esse processo não tenha um bom aproveitamento [Suresh et al., 2009].

Por conta disso, pode-se otimizar vários critérios de validação de grupos para ter certeza de que um agrupamento é realmente bom. Um problema de otimização que trata simultaneamente de vários critérios, ou funções objetivo, é chamado de multiobjetivo e é bastante comum sua ocorrência em problemas da vida real [Suresh et al., 2009]. Por exemplo, no processo de desenvolvimento de *software*, em que a obtenção do sucesso do produto depende, primeiramente, do cumprimento de certas restrições, como não ultrapassar o orçamento planejado, e da otimização de alguns objetivos como a minimização do tempo de desenvolvimento e da quantidade de componentes na equipe desenvolvedora, além da maximização da qualidade do produto final. Em [Freitas, 2004] é citado um exemplo de problema multiobjetivo na etapa de seleção de atributos, em que um subconjunto ótimo de atributos deve maximizar a precisão e minimizar o número de atributos selecionados.

Muitas vezes, como nos exemplos citados, a combinação de várias funções objetivo acarreta uma situação conflitante, em que a otimização de uma função ou objetivo leva à degradação da qualidade de outro objetivo, nos fornecendo desta forma mais de uma solução [Suresh et al., 2009], [Sierra & Coello, 2006].

Em [Suresh et al., 2009] encontra-se uma formulação matemática, apresentada na Equação 4.1, para um problema de otimização multiobjetivo composto por m objetivos, cada um com n parâmetros, também chamados de variáveis de decisão.

$$Otimizar\ \vec{Y}\ = \vec{f}(\vec{Y}) = \left(f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n)\right) \quad (4.1)$$

Com a presença de várias soluções, há a necessidade de selecionar aquelas que apresentam melhores resultados. Na otimização de problemas com múltiplos objetivos há três abordagens que se encarregam da geração e seleção dessas soluções: a formulação baseada em peso, a técnica lexicográfica e a técnica baseada na frente de Pareto [Sierra & Coello, 2006], [Pappa et al., 2004].

Em todas as abordagens, o papel de um usuário com conhecimento no problema, a quem é referenciado como *decision maker*, é de grande importância para fazer algumas.

4.1. Fórmula Baseada em Peso

A técnica da fórmula baseada em peso é a mais simples entre as técnicas de otimização multiobjetivo. Nela, cada objetivo tem um peso associado e determinado pelo tomador de decisão ("*decision maker*") que representa sua importância no resultado e que depois é combinado com os valores dos demais objetivos, de modo a formar um único valor. Dessa forma, os valores de todas as variáveis de decisão são combinadas em uma só fórmula, ou seja, o problema multiobjetivo é transformado em um problema com um único objetivo.

Apesar de ser simples, ela apresenta alguns problemas que estimulam o uso de outras técnicas. Primeiro pode haver dificuldade de ajustar os pesos, já que é raro conhecer a importância de cada objetivo *a priori* e, uma vez feito isso, o algoritmo irá buscar pela melhor solução específica para essa configuração, que não será necessariamente a melhor possível [Freitas, 2004].

O segundo problema consiste em envolver critérios não-comensuráveis na mesma fórmula, em que a combinação de seus pesos pode ser feita através de uma soma ou subtração e assim retornar resultados sem significado algum, além de não considerar situações conflitantes e resultar apenas uma solução.

Por exemplo, se fosse feita uma otimização envolvendo a quantidade de genes selecionados e o índice Jaccard de avaliação de grupos, cujo valor máximo é 1. No método de otimização multiobjetivo baseado na fórmula de peso, os dois critérios se misturariam na mesma fórmula. Como eles possuem grandezas distintas, a proporção da quantidade de genes, que pode ser de milhares, acaba sendo muito maior que a proporção dos valores Jaccard, que

pode ser no máximo 1. O objetivo da otimização é escolher a solução com menor quantidade de genes e maior valor Jaccard. Pode acontecer do método de otimização escolher a solução com uma quantidade muito menor de genes mas com um valor Jaccard não muito bom, pois a combinação de critérios fica desproporcional.

4.2. Técnica Lexicográfica

Uma forma de solucionar o problema de combinar critérios não comensuráveis da técnica da fórmula baseada em peso, é a técnica lexicográfica. Ela trata os critérios de forma independente através de prioridades diferentes para cada objetivo.

Na Expressão 4.1 é apresentada a avaliação funcional lexicográfica, ou LEF, composta por n pares e organizados pela sequência do objetivo mais importante para o objetivo menos importante. Cada par é formado pelo objetivo c e pela sua respectiva tolerância t [Kaufmann & Michalski, 1999].

$$\langle (c_1, t_1), (c_2, t_2), \dots, (c_n, t_n) \rangle \tag{4.1}$$

A tolerância, que é estabelecida pelo tomador de decisão, é a faixa definida entre o melhor valor das variáveis de decisão para um determinado objetivo e o limite dos valores que podem ser considerados aceitáveis. Se uma solução tiver um valor fora da tolerância, é removida do conjunto das melhores soluções.

A avaliação percorre a sequência de prioridade dos objetivos enquanto houver mais de uma solução possível no conjunto das melhores soluções.

Apesar de ser mais complexa que a técnica baseada em pesos, ainda é simples e de fácil uso. Porém, o modo como são aplicadas as prioridades dos critérios acaba limitando em certo grau o espaço de possibilidades a serem exploradas [Freitas, 2004].

4.3. Técnica Baseada na Frente de Pareto

Aqui, o objetivo final é obter um conjunto de soluções ótimas, ao invés de uma única, como nas demais técnicas. Por isso, diz-se que esta técnica resolve os problemas multiobjetivos com métodos multiobjetivos [Freitas, 2004].

Uma solução ótima, ou um vetor de variáveis de decisão ótimo, primeiro tem que ser definida para um problema de Maximização ou Minimização e deve satisfazer a chamada regra de dominância de Pareto.

Dadas duas soluções $\vec{x}$ e $\vec{y}$, pertencentes ao espaço de soluções π, dizemos que $\vec{x}$ domina $\vec{y}$, ou $\vec{x} \prec \vec{y}$, em um problema de minimização com n critérios, se:

- existe pelo menos um critério $i = 1,\ldots,n$ em que x_i é estritamente melhor que y_i, ou seja $x_i \prec y_i$;

- x_i não é pior que y_i em nenhum critério i, ou seja $x_i \preceq y_i$;

Se não existir solução $\vec{x}' \in \pi$, em que $f(\vec{x}') \prec f(\vec{x})$, $\vec{x}$ é Pareto Ótimo. Por sua vez, determinando o conjunto de todos os vetores do Pareto Ótimo, que geralmente consiste em um problema NP-Completo, é possível gerar a frente de Pareto (Figura 4.1). Dentro deste conjunto será selecionada a melhor solução por parte de um agente de decisão, o tomador de decisão, que pode ser o próprio usuário [Coello, Lamont & Vldhuizen, 2007].

Neste trabalho, os conceitos da otimização de problemas multiobjetivos baseados na frente de Pareto serão empregados a seguir na técnica de enxame, um algoritmo evolucionário que, por causa de sua natureza baseada em população, permite a geração de vários elementos do Pareto Ótimo em apenas uma execução [Coello, Lamont & Vldhuizen, 2007].

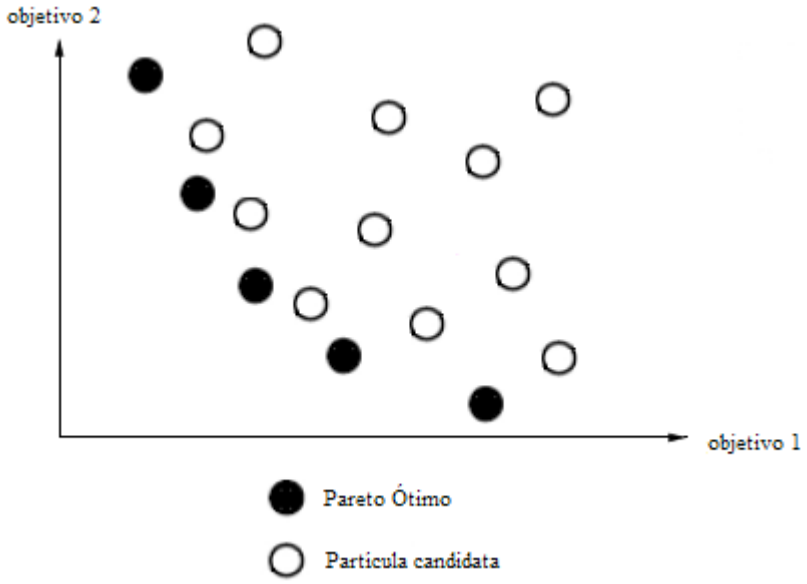

Figura 4.1: Soluções de um problema multiobjetivo

4.4. *Multi-objective Particle Swarm Optimization* (MOPSO)

A técnica de enxame PSO, do inglês *Particle Swarm Optimization,* ou nuvem de partículas, é um algoritmo evolucionário usado tanto para problemas de otimização de um único objetivo quanto para multiobjetivos. Ele se baseia na simulação do movimento de vôo que uma população de pássaros faz em busca de alimentos [Kennedy & Eberhart, 1995].

Sua versão para problemas multiobjetivos, o MOPSO, tornou-se popular por ser simples e pela sua eficiência em varias aplicações, produzindo bons resultados com baixo custo computacional, além da velocidade de convergência em comparação com outras técnicas evolutivas [Chuang et al, 2008].

Cada partícula do enxame representa uma possível solução que se movimenta, ou "voa", pelo espaço de busca. Aquelas que estiverem situadas em regiões promissoras, ou que sejam não-dominadas, são tidas como as melhores soluções e são chamadas de líderes, sendo mais de uma porque se trata de problemas multiobjetivos, mas em que apenas uma é selecionada como líder para cada partícula [Carvalho & Pozo, 2009]. Esses líderes são armazenados em um repositório no qual seu conteúdo é retornado como resultado ao final do algoritmo [Sierra & Coello, 2006].

A direção $\overrightarrow{x_i}(t)$ para onde uma partícula p_i se movimenta em um determinado tempo t, presente na Equação 4.2, é definida pela sua melhor posição $\vec{x}_{pbest_i}$ e pela melhor posição do enxame, que é a posição do líder $\vec{x}_{leader}$ [Carvalho & Pozo, 2009].

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \tag{4.2}$$

Esta equação é influenciada por um operador que simula sua velocidade $\overrightarrow{v_i}(t)$, apresentada na Equação 4.3, que atualiza a posição da partícula corrente e seleciona os líderes ao final de cada iteração, ou tempo. Seu resultado depende do valor da inércia W, que mede o impacto da velocidade anterior sobre a velocidade atual, dos fatores de aprendizagem cognitiva C_1 e de aprendizagem social C_2, que são as influências que uma partícula tem sobre sua própria velocidade e sobre a velocidade da vizinhança, respectivamente, e pelos valores aleatórios $r_1, r_2 \in [0,1]$ [Sierra & Coello, 2006].

$$\vec{v}_i(t) = W\vec{v}_i(t-1) + C_1 r_1 \left(\vec{x}_{pbest_i} - \vec{x}_i(t-1)\right) + C_2 r_2 \left(\vec{x}_{leader} - \vec{x}_i(t-1)\right) \tag{4.3}$$

No PSO, dois fatores podem afetar diretamente o resultado final: a topologia usada para indicar a vizinhança de uma partícula e o método para selecionar o líder de cada partícula.

A forma como as partículas estão interligadas, ou seja, o tipo de topologia usada para indicar a influência de uma partícula sobre as outras vai determinar a vizinhança de uma partícula, sendo que topologias diferentes constroem vizinhanças diferentes, e dentro desse conjunto de vizinhos será selecionado um para exercer a função de líder, que pode ser escolhido pelo método Sigma que será visto mais adiante [Sierra & Coello, 2006].

Na topologia grafo vazio, cada partícula é independente das demais. Portanto, no cálculo de seu posicionamento, C_2 é nulo. Na topologia chamada melhor local, uma partícula está ligada a k partículas vizinhas e aquela que for considerada vizinho mais próximo será o líder para esta partícula [Carvalho & Pozo, 2009]. Nos casos em que $k = 2$, tem-se a topologia em forma de anel (Figura 4.2a), e se todas as partículas estiverem interligadas, formará um grafo completamente conectado, conforme Figura 4.2b. Quando somente uma partícula está ligada com as demais e essas não têm outra ligação, é formada a topologia em forma de estrela (Figura 4.2c). Ao ponto central dessa formação é dado o nome de partícula focal, que também se torna líder do enxame [Sierra & Coello, 2006]. Por último, tem-se a topologia em forma de árvore (Figura 4.2d), em que cada partícula é influenciada pela sua posição e pela posição de seu pai na árvore. Caso um nó filho seja uma solução melhor que seu nó pai, há uma troca de posição entre eles. Esta organização hierárquica torna a busca mais dinâmica, mais rápida, porém mais complexa [Carvalho & Pozo, 2009].

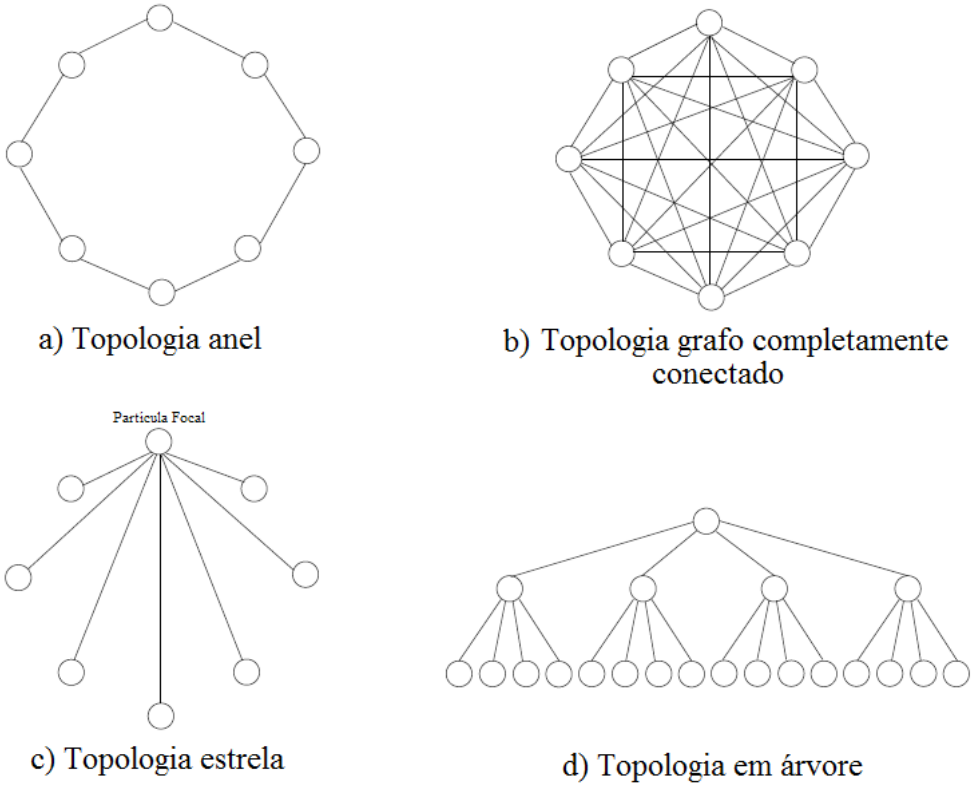

Figura 4.2: Topologias do PSO [Sierra & Coello, 2006]

Uma vez determinada a topologia de vizinhança, outro método é usado para selecionar os líderes, e cada método seleciona líderes diferentes. Em [Mostaghim & Teich, 2003] é apresentado o método sigma, que procura retornar soluções boas e diversificadas em uma quantidade menor de iterações [Carvalho & Pozo, 2009], tornando-se um método muito eficaz em problemas multiobjetivos baseados em técnicas de enxame [Sierra & Coello, 2006].

O vetor $\vec{\sigma}$, calculado pela Equação 4.4, realiza a combinação dos objetivos e é composto por $\binom{m}{2}$ elementos, em que m é a dimensão do espaço objetivo. O líder cujo vetor sigma é mais próximo que o vetor sigma da partícula, e que pertence à mesma vizinhança, será eleito guia para essa partícula [Mostaghim & Teich 2003].

$$\vec{\sigma} = \begin{pmatrix} {f_1}^2 - {f_2}^2 \\ {f_1}^2 - {f_3}^2 \\ \cdots \\ {f_{m-1}}^2 - {f_m}^2 \end{pmatrix} \Big/ \left({f_1}^2 + {f_2}^2 + {f_3}^2 + \cdots + {f_{m-1}}^2 + {f_m}^2\right) \qquad (4.4)$$

A seleção do líder para uma partícula é feita primeiramente pelo cálculo do vetor sigma da própria partícula $\vec{\sigma}_p$ e de cada solução i não-dominada do repositório $\vec{\sigma}_{leader_i}$. O

líder cuja distância do seu sigma para o sigma da partícula da população for menor, conforme Figura 4.3, será eleito líder para esta partícula [Mostaghim & Teich, 2003].

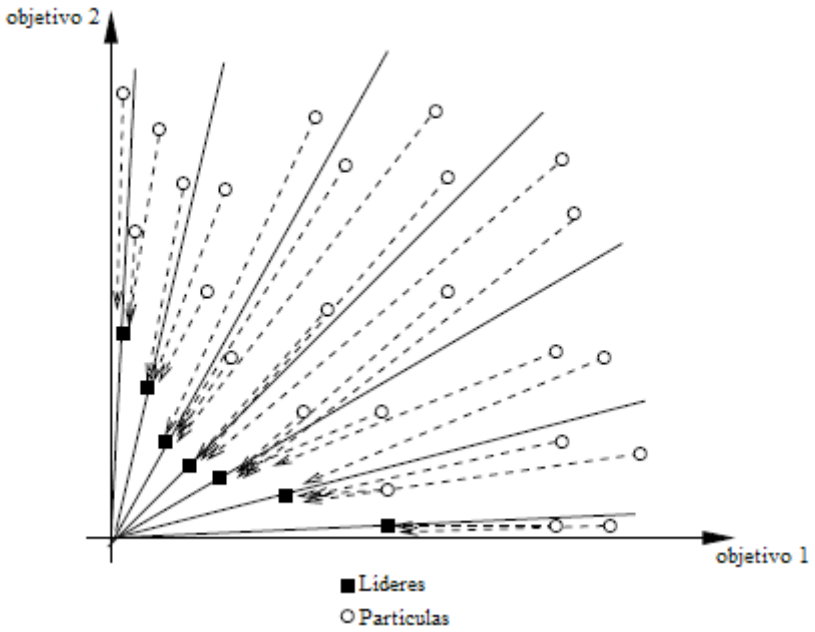

Figura 4.3: Escolha de líderes pelo método Sigma [Mostaghim & Teich, 2003]

De uma forma geral, o MOPSO neste trabalho funciona da seguinte maneira (Algoritmo 3.1). Primeiramente o enxame é inicializado com as soluções obtidas pelos critérios de avaliação de grupos e, pela regra de dominância de Pareto, são eleitos os líderes do enxame e armazenados em um repositório L.

Em todas as iterações, cada partícula executa seu vôo em relação ao seu guia que é escolhido pelo método sigma e atualiza a sua melhor posição S_{best} já alcançada. Quando todas as partículas realizarem seus vôos, é realizada a atualização dos líderes no repositório independente. Ao término do algoritmo, quando o número máximo de iterações for alcançado ou quando não houver modificações no repositório de líderes, o conteúdo desse repositório é retornado com as melhores soluções.

Algoritmo 3.1 PSO

```
Inicializa o Enxame
Elege líderes e inicializa o repositório L
Iteração = 0 //Número da iteração corrente
Enquanto critério de parada não alcançado
```

```
    Para cada partícula
        Seleciona o líder em L
        Realiza o vôo
        Atualiza o S_best
    Fim para
    Atualiza L
    Iteração++
Fim enquanto
Retorna L
```

4.5. Considerações Finais

Neste capítulo foram apresentadas técnicas multiobjetivo, que otimizam simultaneamente vários objetivos e são usadas para otimizar vários índices de validação de grupos.

Há 3 abordagens de técnicas multiobjetivo: a fórmula baseada em peso, a técnica lexicográfica e a técnica baseada na frente de Pareto. Este último é o único cuja influência do tomador de decisão ocorre somente no fim da execução, na escolha da melhor solução e não durante o desenvolvimento.

A técnica baseada em enxame PSO é uma aplicação da frente de Pareto e sua versão multiobjetivo, o MOPSO, pode selecionar várias soluções tidas como ótimas.

Cada solução usada nas técnicas multiobjetivo neste trabalho é composta por valores dos índices de validação de grupos, que representam diferentes métodos de seleção de atributos. Logo, o uso do MOPSO servirá para escolher as melhores soluções para uma determinada característica, avaliando assim os métodos de seleção de atributos.

Capítulo 5

Projeto

Este trabalho tem como objetivo principal avaliar, por meio de uma otimização multiobjetivo, métodos de seleção de atributos em bases de dados obtidas pela técnica do microarranjo. Como resultado, espera-se selecionar os métodos de seleção ótimos e identificar a quantidade dos atributos mais relevantes para uma característica, a fim de facilitar a análise da base de dados obtidas a partir de microarranjos, que são formadas por milhares de atributos.

A seguir, neste capítulo, serão vistas as etapas realizadas e as bases de dados de expressões gênicas analisadas.

5.1. Etapas do projeto

As etapas realizadas neste trabalho são mostradas na Figura 5.1. Nela é ilustrada a redução de dimensionalidade da base de dados original por apenas um método de seleção de atributos. A qualidade desse método é medida pelo uso de m métodos de agrupamento e, para cada agrupamento, n índices de validação de grupos. As soluções formadas por vetores contendo os valores dos índices de validação de grupos são submetidas a otimização multiobjetivo, quando será escolhido o melhor método de seleção de atributos e, consequentemente, os atributos mais relevantes.

Tem-se, como pressuposto inicial, a disponibilidade e o conhecimento das características de bases de dados originais de microarranjos. Elas consistem em descrever o comportamento dos genes em relação a doenças cancerígenas, como o Linfoma e a Leucemia, sob diversas condições, com a finalidade de selecionar aqueles que são mais relevantes para tais doenças.

Essas mesmas bases também já disponibilizam uma descrição das condições em que as mesmas foram obtidas e todas as características relativas às mesmas são descritas em detalhes mais adiante neste capítulo. Alguns conceitos básicos sobre as doenças envolvidas e uma breve descrição das bases serão vistas mais adiante, ainda neste capítulo, para um melhor esclarecimento sobre o ambiente proposto.

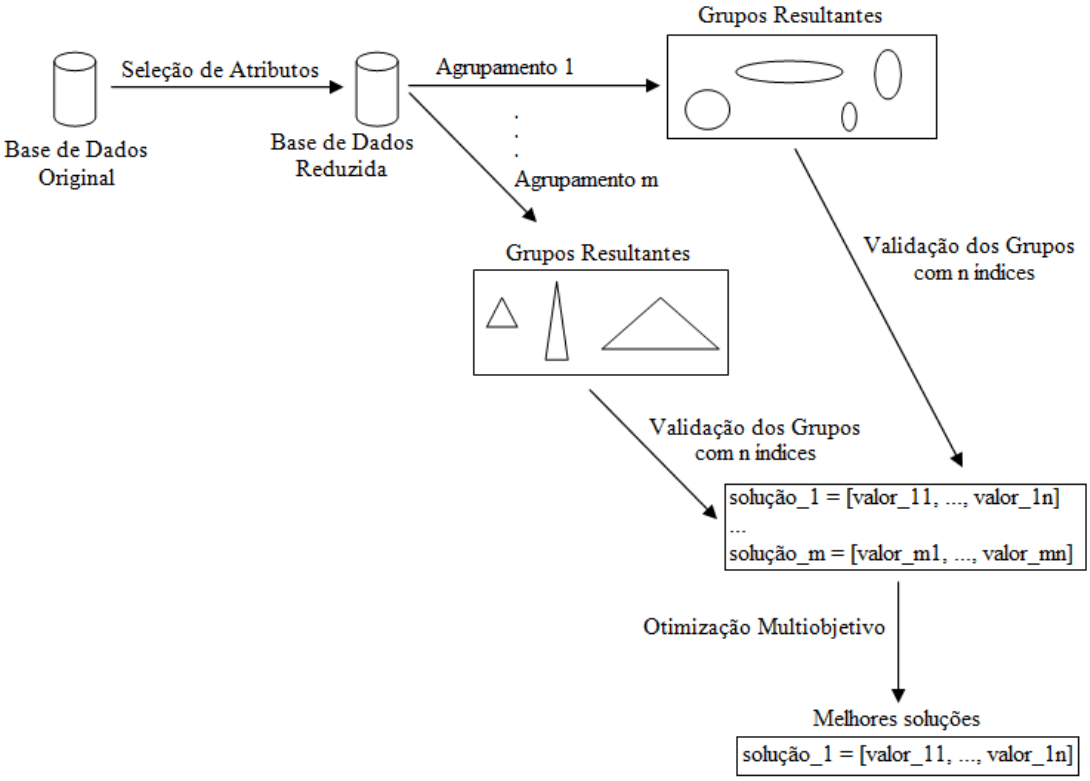

Figura 5.1: Etapas do projeto

Pelo fato de todas as bases estarem disponíveis no formato "arff", as análises realizadas neste trabalho foram feitas utilizando o software Weka[1], que contém um conjunto de bibliotecas relacionadas a mineração de dados.

Essas bases foram submetidas ao processo de seleção de atributos por três métodos: C-*FOCUS*, CFS e o *Relief*-F, esse último reduzindo cada base a 10%, 25%, 50% e 75% da quantidade de seus atributos originais. Os demais parâmetros dos métodos de seleção de atributos usaram valores padrão do Weka e ao final desse processo existirão, para cada base original, seis novas bases reduzidas.

[1] http://www.cs.waikato.ac.nz/ml/weka/, acessado em 07 de Abril de 2011.

Cada base reduzida pela seleção de atributos, assim como as bases originais, será qualificada através do agrupamento das instâncias. Neste trabalho serão abordados três métodos de agrupamento pertencentes a diferentes abordagens. Primeiramente o K-*means*, que é baseado em protótipo e exige que o usuário forneça a quantidade de grupos a serem usados. Para realizar um processo de agrupamento com menos influência do usuário, é usado o método ISODATA, uma evolução do K-*means*.

O terceiro método de agrupamento utilizado é o Classit, que é hierárquico conceitual e usado de forma padrão estabelecida pelo Weka. Seu ponto de corte é especificado no primeiro nível a partir da raiz, onde há uma generalização maior dos grupos.

Por sua vez, a qualidade do agrupamento é representada pelo vetor contendo os valores dos índices de validação de grupos. Também serão realizados experimentos envolvendo diferentes grupos de índices para observar há comportamentos diferentes ou se eles trabalham de forma semelhante.

Para os índices relativos (Dunn, C, Davies-Bouldin e *Silhouette*), o parâmetro da quantidade de grupos é instanciado pelo método de agrupamento que eles vão qualificar. O índice interno *Isolation* usa apenas informações pertencentes às bases de dados. Já para o índice externo Jaccard, por poder comparar dois agrupamentos, serão comparados o agrupamento resultante por cada método de agrupamento e seu respectivo agrupamento real, já que há um conhecimento *a priori* sobre a composição dos grupos das bases de dados.

Ao final do processo de agrupamento serão formados, conforme Equação 5.1, x vetores de índices.

$$n° \, métodos \, de \, seleção \times n° \, métodos \, de \, agrupamento = x \, soluções \quad (5.1)$$

A otimização desses vetores identificará o melhor método de seleção de atributos. Pelo fato de serem usados vários critérios, é caracterizado um problema multiobjetivo. Primeiramente será usado o método MOPSO para revelar os melhores métodos de seleção.

Os resultados obtidos pelo MOPSO serão comparados com resultados obtidos por mais outros dois métodos multiobjetivo: o método lexicográfico e o método baseado em fórmula de pesos.

Este trabalho faz uso do funcionamento evolucionário do método MOPSO baseado na frente de Pareto para movimentar as soluções, ou partículas, seguindo seus respectivos guias

que foram escolhidos pelo método sigma para as regiões promissoras. Os líderes eleitos pela regra de dominância de Pareto ao final da execução, quando não houver alteração no repositório, serão considerados soluções ótimas, revelando os melhores métodos de seleção de atributos para esses critérios.

Por se tratar de poucas soluções, todas as partículas fazem parte da mesma vizinhança, formando um grafo completamente conectado. Segundo [Sierra & Coello 2006], essa topologia converge para o resultado final mais rapidamente.

No cálculo da velocidade de cada partícula, $r_1, r_2 \in [0,1]$ foram escolhidos aleatoriamente a cada iteração. O valor da inércia foi fixada em 1, que é considerado um valor alto e ajuda na exploração global, já que todas as partículas fazem parte da mesma vizinhança [Sierra & Coello 2006]. As constantes de aprendizagem também foram fixadas e o valor 2 foi escolhido a partir de experiência adquirida por [Eberhart & Shi, 2001].

O MOPSO objetiva otimizar o enxame através da maximização dos índices Dunn, Silhueta e Jaccard, e através da minimização dos índices C, Davies-Bouldin e *Isolation.*

Para ilustra o funcionamento do MOPSO, serão utilizadas três partículas (Figura 5.2). Elas representam soluções obtidas por diferentes métodos de seleção e agrupamento para uma mesma base de dados. A movimentação e escolha da melhor partícula é feita em relação ao índice Jaccard e ao número de atributos que foram selecionados.

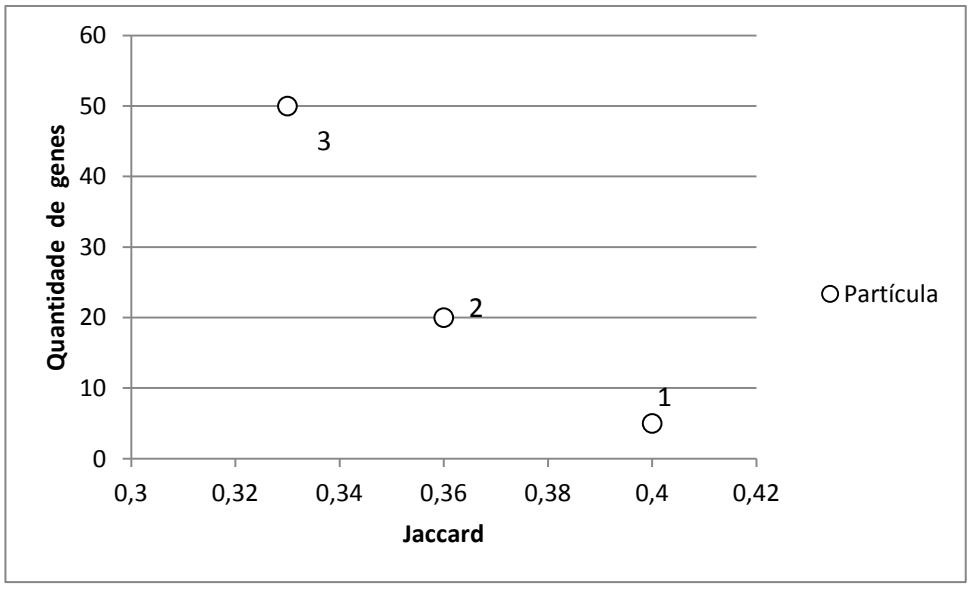

Figura 5.2: Posição das partículas

Como pode ser vista na Figura 5.2, a partícula 1 otimiza tanto o índice Jaccard, por ter o maior valor, quanto o número de genes selecionados. Pela regra de dominância de Pareto, esta solução domina as demais e é eleita líder da vizinhança. Dessa forma, as soluções 2 e 3 irão se deslocar em direção a solução 1. Por se tratar de um problema multiobjetivo, é possível ter mais de um líder, caso ele não seja dominado por ninguém.

Após o cálculo da velocidade e a atualização da posição das partículas, um novo cenário pode ser visto na Figura 5.3.

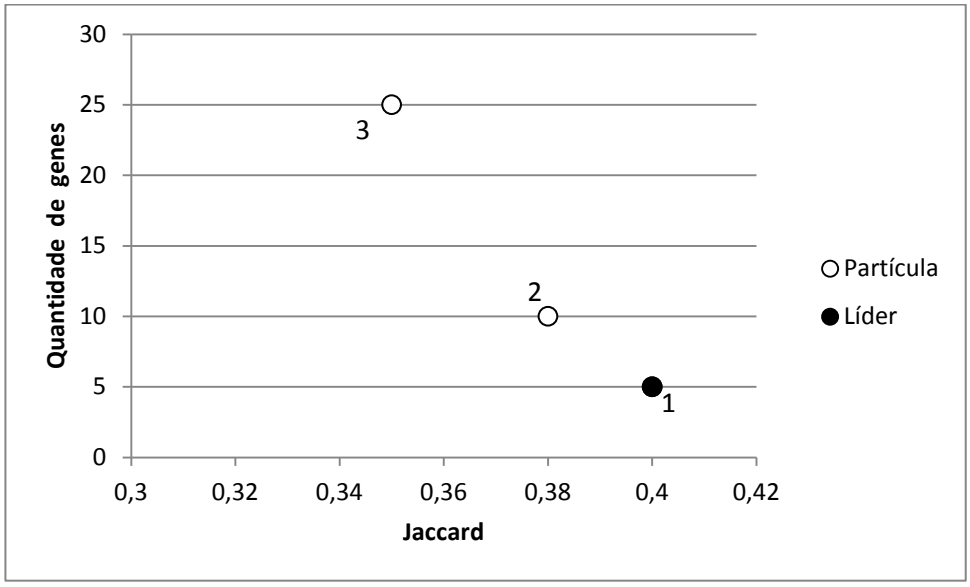

Figura 5.3: Posição das partículas pós-MOPSO

As partículas 2 e 3 estão mais próximas do líder, a partícula 1. Pode ocorrer de uma partícula se tornar melhor que o presente líder, havendo assim uma troca de liderança. No exemplo apresentado, como não houve essa mudança, o processamento é encerrado e a partícula 1 é eleita a melhor solução e o método de seleção a qual ele representa é dito melhor para a base utilizada.

No método lexicográfico, a sequência de importância de objetivos varia de acordo com o experimento mas em geral é composta, do melhor para o pior, pelos índices externos, relativos e, por último, internos.

Primeiramente são usados os índices externos pelo fato de comparar o agrupamento obtido por cada método pelo agrupamento real. Em seguida são usados os índices relativos, já que há conhecimento *a priori* sobre a quantidade exata de grupos nas bases. Em caso de não haver este conhecimento, métodos de agrupamento capazes de gerar a quantidade correta de

grupos deverão ser utilizados. Por último, os índices internos são usados. Dentro dos índices relativos, a sequência é definida, do melhor para o pior, por: C, Silhueta, Dunn e Davies-bouldin.

Todos os objetivos usam como valor de tolerância 30% acima do melhor valor, para critérios de minimização, e 30% abaixo do melhor valor para critérios de maximização. Tanto a sequência de preferência de critérios, quanto o valor da tolerância foram escolhidos após a realização de vários experimentos. Valores de tolerância maiores elegeram precocemente as melhores soluções, sem levar em consideração a maioria dos critérios. Já valores de tolerância menores não foram capazes de eliminar coerentemente as piores soluções.

Para o método baseado na fórmula de pesos não há uma sequência pré-estabelecida e todos os objetivos são combinados na mesma fórmula. Foi definido que todos os valores dos objetivos serão somados, sendo que aqueles cuja meta de otimização seja a maximização tenham seus valores multiplicados pelo valor 1, que é o peso deles. Para os objetivos que visam minimizar seus valores são atribuídos peso -1.

Os valores de cada objetivo das soluções foram divididos pelo melhor valor encontrado. Assim, não haverá grande diferença entre os objetivos que têm diferentes níveis de valores. A definição desses pesos deu-se pelo fato de ser uma abordagem simples e que obteve bons resultados.

Todos os experimentos foram realizados em uma máquina Intel ® Core™ I7 com 1.73GHz e 6GB de memória RAM. O MOPSO baseado na frente de Pareto, o método lexicográfico e o método baseado na fórmula de pesos foram desenvolvidos na linguagem Java, assim como os métodos de seleção de atributos, métodos de agrupamento e os índices de validação de grupos, que utilizaram a biblioteca do *software* Weka.

5.2. Bases de Dados

As bases de dados de expressões gênicas que foram usadas nos experimentos que serão explicados nesta sub-seção foram extraídas da técnica de microarranjo. Eles têm como tema quatro tipos de cânceres: o Linfoma Difuso de Grandes Células B (LDGCB ou DLBCL, do inglês), o Linfoma Folicular (LF ou FL) e as Leucemias Agudas Mielóide (LAM ou AML) e Linfoblástica (LAL ou LLA).

A seguir serão apresentadas as cinco bases de dados usadas neste trabalho (Tabela 5.1). Elas, que também foram utilizadas em [Borges, 2006], estão disponíveis, assim como

documentos que referenciam essas bases, na página da Kent Ridge[2] em formato "arff", que é o formato utilizado no *software* Weka.

5.2.1. DLBCL - Stanford

Através do mapeamento genético no micro-arranjo Lymphochip, foi descrito o comportamento dos genes em diferentes fases no processo de ativação da célula B para descobrir padrões que são suficientes para a origem do Linfoma Difuso de Grandes Células B. [Alizadeh et al., 2000] observaram que, na fase em que o linfócito passa pelo centro germinativo do folículo linfático, uma grande quantidade de genes se comportam de forma suspeita originar esta doença. Outra fase observada a fim de observar as expressões gênicas para este linfoma é no momento de ativação do linfócito B.

Esta base é composta por 47 amostras: 24 pertencentes ao grupo do centro germinativo e 23 pertencentes ao grupo de ativação da célula B. Cada amostra é descrita por 4026 genes.

5.2.2. DLBCL-Tumor

Nesta base de dados está descrito o comportamento dos genes para dois tipos de Linfoma: o LDGCB e o LF. Outro estudo fazendo comparações entre essas duas doenças e utilizando a mesma base foi feita em [Shipp et al., 2002].

O conjunto de dados é composto por 77 amostras: 58 pertencentes aos pacientes do LDGCB e 19 pertencentes aos pacientes do LF. Cada amostra é descrita por 6817 genes.

5.2.3. DLBCL - Outcome

Esta base, também estudada em [Shipp et al., 2002], descreve o comportamento dos genes para os pacientes do LDGCB após 5 anos de tratamento, mostrando que esse tempo é um fator que indica, geralmente, a cura de uma pessoa com esse Linfoma.

O conjunto de dados é composto por 58 amostras: 32 pertencentes aos pacientes do LDGCB considerados curados e 26 pertencentes aos pacientes que sofreram fatalidades ou cuja doença ainda persiste. Cada amostra é descrita por 6817 genes.

5.2.4. DLBCL - NIH

A sobrevivência de pacientes com LDGCB após a quimioterapia é influenciada por características moleculares dos tumores [Rosenwald et al., 2002]. Esta base de dados ajuda na

[2] http://datam.i2r.a-star.edu.sg/datasets/krbd/, acessado em 18 de Julho de 2011.

visualização de padrões quanto ao comportamento de 7399 genes de 80 pacientes, após a terapia, para serem feitas predições quanto às condições do paciente. Entre os pacientes, 30 deles foram rotulados vivos e o restante falecidos.

5.2.5. Leukemia-ALL/AML

O tratamento para os vários tipos de leucemia, assim como os diversos tipos de cânceres, nem sempre é o mesmo. Por isso, é importante fazer o diagnóstico correto da doença para que seja providenciado o tratamento correto. Esta base de dados, que também foi usada em [Golub et al., 1999], descreve o comportamento dos genes para dois tipos de Leucemia, a ALL e a AML.

O conjunto de dados da base é formado por 34 amostras: 20 pertencentes aos pacientes da doença ALL e 14 pertencentes aos portadores da AML. Cada amostra é descrita por 6817 genes.

5.2.6. Leukemia - MLL

Quando o paciente com leucemia aguda é uma criança, uma translocação envolvendo o gene MLL (*Mixed-Lineage Leukemia*) pode ocorrer no tratamento quimioterápico, caracterizando a Leucemia MLL. Quando isso ocorre, a criança pode sofrer uma recaída após a quimioterapia e novas recomendações no tratamento devem ser feitas, diferenciando-se do tratamento imposto aos pacientes com ALL e AML [Armstrong et al., 2002].

O conjunto de dados da base é formado por 57 amostras: 20 pertencentes aos pacientes com ALL, 17 pertencentes aos portadores de MLL e outros 20 com AML. Cada amostra é descrita por 12582 genes.

Tabela 5.1: Descrição das bases de dados

Nome	Atributos	Instâncias	Grupo 1	Grupo 2	Grupo 3
DLBCL - Stanford	4026	47	24	23	0
DLBCL - Tumor	7129	77	58	19	0
DLBCL - Outcome	7129	58	32	26	0
DLBCL – NIH	7399	80	30	50	0
Leukemia-ALL/AML	7129	34	20	14	0
Leukemia-MLL	12582	57	20	17	20

5.3. Trabalhos Relacionados

Nesta seção serão abordados alguns trabalhos que envolveram redução de dimensionalidade ou o uso do MOPSO para otimizar problemas em diferentes áreas. Esses trabalhos serviram como contribuições para o desenvolvimento do presente trabalho.

Em [Borges, 2006], a redução de dimensionalidade é usada na tarefa de classificar dados de microarranjos. São utilizadas a seleção de atributos e o método de projeção aleatória em classificadores como o SVM e o Naive Bayes. Os métodos de seleção de atributos obtiveram melhores resultados comparados aos métodos de projeção e se mostraram alternativas na mineração de dados de expressão gênica, em que os métodos *wrapper* exigiram mais computacionalmente.

O MOPSO também tem sido tema de pesquisas em diferentes áreas, como em [Carvalho & Pozo, 2009]. Nele, o MOPSO busca avaliar problemas de aprendizado em regras na mineração de dados e conseguiu produzir bons conjuntos de regras, além de ser comparado com outros algoritmos de aprendizado de regras.

O método proposto neste trabalho vem evoluindo desde [Garcia & Nievola, 2011], quando o MOPSO otimizou o índice Jaccard de validação de grupos e a quantidade de atributos selecionados pelos métodos de redução de dimensionalidade. Apesar de o MOPSO classificar como melhores as soluções pertencentes às menores bases, concluiu-se que a quantidade de atributos não é um bom objetivo ao ser otimizado junto a um índice vindo do agrupamento. Isso porque a diferença da quantidade de atributos entre as bases originais e as bases reduzidas é muito grande, tornando o trabalho do MOPSO não necessário.

Já em [Garcia, Paraiso & Nievola, 2011] foram utilizados índices de validação de agrupamento pertencentes a diferentes tipo, como relativo (índice C), interno (índice *isolation*) e externo (Jaccard). Dessa forma, a avaliação do MOPSo se tornou mais coerente.

5.4. Considerações Finais

Nesse capítulo foram apresentadas as etapas necessárias para realizar o processo de avaliação dos métodos de seleção de atributos em bases de dados de expressões gênicas obtidos pela técnica do microarranjo. Essa avaliação é feita através da qualidade dos agrupamentos das instâncias das bases que são medidas pelos índices de validação de grupos.

O uso de vários índices caracteriza um problema de otimização multiobjetivo. Logo, métodos multiobjetivo são usados para otimizar simultaneamente esses índices.

As descrições dos experimentos, assim como os resultados que foram obtidos através da realização das etapas da proposta, serão vistas no capítulo a seguir.

Capítulo 6

Experimentos e Resultados

Neste capítulo serão apresentados os experimentos e os resultados mais importantes, já que há muitos resultados que não acrescentam comportamentos especiais. Todos os resultados poderão ser vistos no Apêndice A.

Cada experimento é composto pelas etapas já apresentadas no capítulo anterior e objetiva avaliar métodos de seleção de atributos em bases de dados de expressões gênicas.

Cada base de dados em questão é analisada em três fases, envolvendo separadamente cada método de agrupamento.

Para cada agrupamento serão realizadas duas etapas: utilizando apenas dois índices de validação de grupos (o índice interno *Isolation* e o índice externo Jaccard), de forma a possibilitar uma melhor visualização do impacto do trabalho exercido pelo MOPSO através de gráficos que mostram as posições das soluções após a otimização multiobjetivo. A segunda etapa corresponde ao uso de todos os índices relativos juntos.

Essa organização de fases é importante para observar separadamente o comportamento dos critérios de validação de grupos e verificar o tipo de agrupamento mais adequado para cada base de dados.

Os resultados apresentados nesta seção são mostrados em tabelas contendo a identificação do método de seleção e os valores dos índices de validação de grupos, sendo que os valores em negrito representam os melhores valores para um índice. Nos experimentos que envolvem somente dois critérios, os resultados pós-MOPSO serão mostrados em gráficos. Nesses gráficos, os líderes escolhidos pela frente de Pareto ao final do MOPSO estão representados pelas bolas pretas. As demais soluções, chamadas de partículas, estão representadas por bolas brancas.

A identificação dos métodos de seleção usados para cada solução está numerada, tanto nas tabelas quanto nas figuras, da seguinte forma:

1- Solução obtida pelo método C-FOCUS;
2- Solução obtida pelo método CFS;
3- Solução obtida pelo método *Relief*-F com 10% do tamanho original;
4- Solução obtida pelo método *Relief*-F com 25% do tamanho original;
5- Solução obtida pelo método *Relief*-F com 50% do tamanho original;
6- Solução obtida pelo método *Relief*-F com 75% do tamanho original;
7- Solução obtida pela base original;

Após a apresentação dos resultados, será mostrado um resumo com a quantidade de vezes que um método de seleção foi escolhido como mais adequado para uma base de dados de expressões gênicas obtida pela técnica do microarranjo. Informações adicionais sobre os atributos que foram mais selecionados poderão ser encontradas no Apêndice B.

6.1. Resultados da base DLBCL-Stanford

O primeiro exemplo de resultado a ser explicado ocorreu na base DLBCL-Stanford. Abaixo, na Tabela 6.1, encontra-se a quantidade de atributos selecionados e seus respectivos métodos de seleção.

Tabela 6.1: Quantidade de atributos selecionados para DLBCL-Stanford

Número do método	Quantidade de atributos
1	3
2	47
3	402
4	1006
5	2013
6	3020
7	4026

Ao agrupar as bases DLBCL-Stanford usando o K-*means*, pode-se observar o comportamento mais encontrado nos experimentos, que é a escolha das menores bases pelos métodos de otimização multiobjetivo.

Primeiramente serão vistos, na Tabela 6.2, os valores dos índices *Isolation* e Jaccard para cada solução. Como já dito, o uso de dois critérios possibilita a visualização de gráficos que mostram como cada partícula, ou cada solução, movimentou no espaço de soluções.

Tabela 6.2: Soluções DLBCL-Stanford usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**1.52**	0.52
2	4.34	**0.77**
3	12.34	0.64
4	17.93	0.54
5	21.18	0.32
6	24.62	0.26
7	27.4	0.26

Para o índice *Isolation*, o objetivo é minimizar seu valor. Logo, a solução 1 obteve o melhor valor. Já para o índice Jaccard, cujo objetivo é maximizar seu valor, a solução 2 foi a melhor.

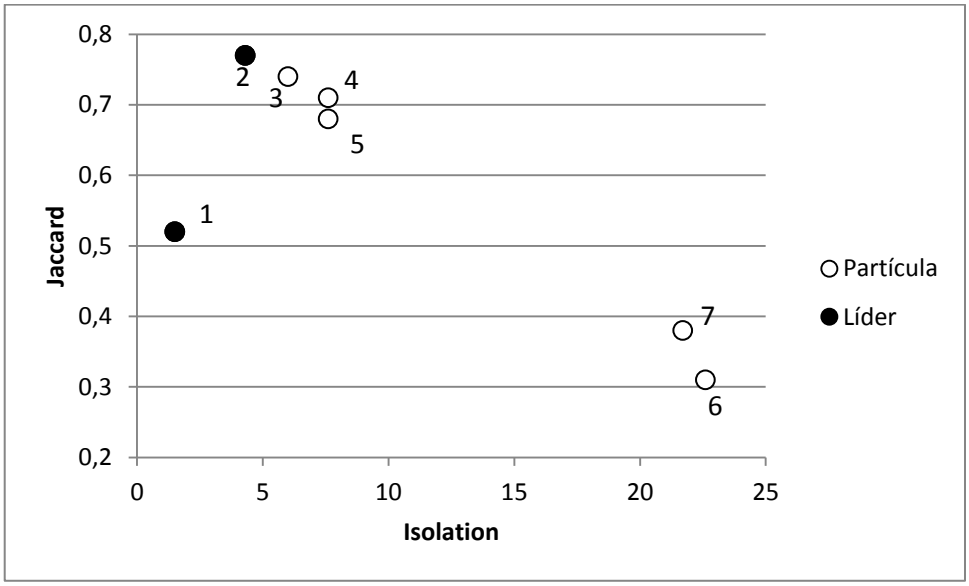

Figura 6.1: Soluções DLBCL-Stanford pós-MOPSO usando K-means, Isolation e Jaccard

Após a execução do MOPSO, na Figura 6.1, as soluções tidas como melhores e que foram chamadas de líderes continuaram sendo as de número 1 e 2. Observa-se que as soluções

3, 4 e 5 se aproximaram da solução 2, que foi eleita o líder para elas pelo método sigma. O mesmo aconteceu para as soluções 6 e 7, que se aproximaram da solução 1.

Utilizando os critérios relativos, é possível descobrir como as soluções se comportaram diante da otimização de diversos objetivos. Na Tabela 6.3 são encontrados os valores dos critérios relativos de agrupamentos realizados pelo K-*means*, onde a solução 2 domina o índice Dunn e a solução 1 domina os índices Davies-Bouldin, C e Silhueta. Desse modo, essas duas soluções foram inicialmente escolhidas como líderes.

Tabela 6.3: Soluções DLBCL-Stanford usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	**1.46**	**1.91**	**0.44**
2	**0.41**	1.63	3.54	0.36
3	0.39	1.97	5.0	0.29
4	0.4	2.61	5.76	0.2
5	0.38	3.0	6.0	0.15
6	0.36	2.99	6.31	0.12
7	0.37	3.14	6.15	0.13

Ao final do MOPSO, os líderes foram mantidos e mais uma vez as partículas se aproximaram deles, como pode ser visto na Tabela 6.4.

Tabela 6.4: Soluções DLBCL-Stanford pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	**1.46**	**1.91**	**0.44**
2	**0.41**	1.63	3.54	0.36
3	0.4	1.89	4.69	0.3
4	0.4	2.38	5.25	0.23
5	0.4	2.13	4.46	0.28
6	0.37	2.77	5.87	0.16
7	0.38	2.48	5.02	0.23

O fato de não haver mudanças de lideranças se deve a quantidade máxima de iterações estabelecida ao MOPSO, pois como as partículas se aproximam dos líderes, se houver uma grande quantidade de iterações, essas partículas vão acabar se deslocando para regiões melhores e se tornando líder.

Neste trabalho, o critério de parada foi estabelecido quando o repositório de líderes não for alterado, certificando que as soluções ditas melhores passem pelo menos duas iterações como líderes e garantindo a escolha por aquelas que realmente obtiveram os melhores valores.

Para os demais métodos de otimização multiobjetivo, o método lexicográfico selecionou a solução do método C-*FOCUS* como a melhor, já que obteve um valor melhor para o índice C além da tolerância. Já o método baseado em fórmula de peso selecionou a solução 2 como melhor.

6.2. Resultados da base DLBCL-Tumor

O segundo exemplo relevante, retirado dos resultados presentes no Apêndice A, compara o comportamento das bases DLBCL-Tumor usando os três métodos de agrupamento: K-*means*, ISODATA e Classit.

Na Tabela 6.5 estão os valores obtidos pelos critérios *Isolation* e Jaccard para as bases em questão usando o K-*means*.

Tabela 6.5: Soluções DLBCL-Tumor usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.43**	0.48
2	1.28	**0.78**
3	4.70	0.42
4	6.77	0.39
5	8.62	0.40
6	6.78	0.45
7	7.42	0.44

Mais uma vez, as menores bases obtiveram os melhores resultados. A solução 1 dominou as demais solução no índice *Isolation*, enquanto que a solução 2 foi a melhor no índice Jaccard. Dessa forma, o MOPSO selecionou essas duas soluções como as melhores, como pode ser visto na Figura 6.2.

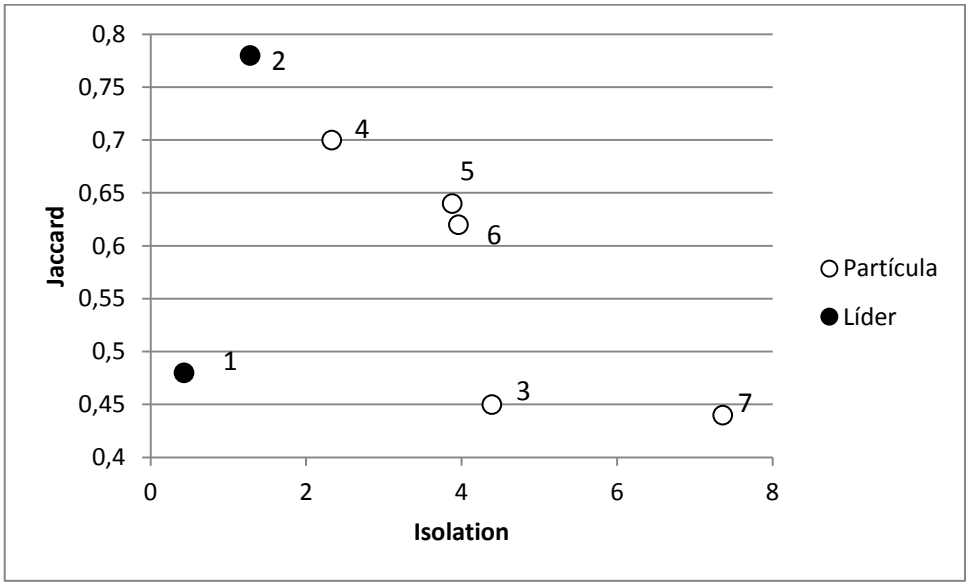

Figura 6.2: Soluções DLBCL-Tumor pós-MOPSO usando K-means, Isolation e Jaccard

O mesmo aconteceu quando as bases DLBCL-Tumor foram agrupadas pelo ISODATA. A seguir, encontra-se na Tabela 6.6 os valores dos índices *Isolation* e Jaccard e depois, na Figura 6.3, as posições finais das soluções após a execução do MOPSO.

Tabela 6.6: Soluções DLBCL-Tumor usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.23**	0.39
2	1.15	**0.54**
3	4.51	0.41
4	7.20	0.40
5	8.50	0.39
6	9.43	0.39
7	10.69	0.46

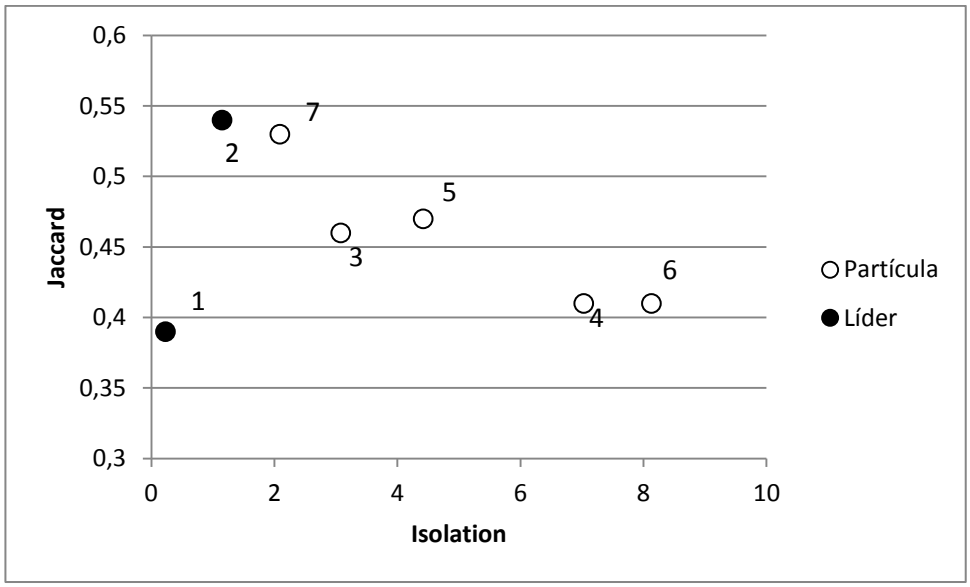

Figura 6.3: Soluções DLBCL-Tumor pós-MOPSO usando ISODATA, Isolation e Jaccard

Pelos resultados observa-se que o ISODATA gerou alguns agrupamentos parecidos com os gerados pelo K-*means*, já que os valores dos índices de validação de grupos são próximos. Além disso, ambos tiveram as mesmas soluções escolhidas como melhores pelo MOPSO.

Por esses fatores, o ISODATA mostra-se uma alternativa ao uso do K-*means*. Além do mais, com o ISODATA não é necessário saber *a priori* o número correto de grupos em um agrupamento, pois este é descoberto pelo auto-ajuste dos grupos durante sua execução.

Tabela 6.7: Soluções DLBCL-Tumor usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	4.42	**0.45**
2	3.50	0.43
3	4.14	0.42
4	**3.36**	0.44
5	3.46	0.44
6	4.59	0.44
7	4.21	0.46

Com relação ao Classit, os resultados na Tabela 6.7 diferiram bastante dos já apresentados para o DLBCL-Tumor, em que a solução 4 otimizou o índice *Isolation* e a solução 7 otimizou o índice Jaccard.

Essa diferença de resultados mostra a importância de utilizar diversos métodos de agrupamento pertencentes a diferentes paradigmas, já que gera diferentes tipos de grupos. É interessante observar a composição dos grupos que a maioria dos métodos gera.

Dessa forma, pode-se fazer uma avaliação mais precisa do formato de grupo presente na base de dados.

Na Figura 6.4 está presente a localização das partículas após a execução do MOPSO. Além de serem eleitas líderes soluções diferentes dos demais experimentos nas bases DLBCL-Tumor, nota-se que uma solução foi adicionada no conjunto de líderes, a solução 3. Esse fato mostra que, a depender da posição inicial das partículas, soluções podem migrar para a região promissora e pode-se haver mudanças no repositório de líderes.

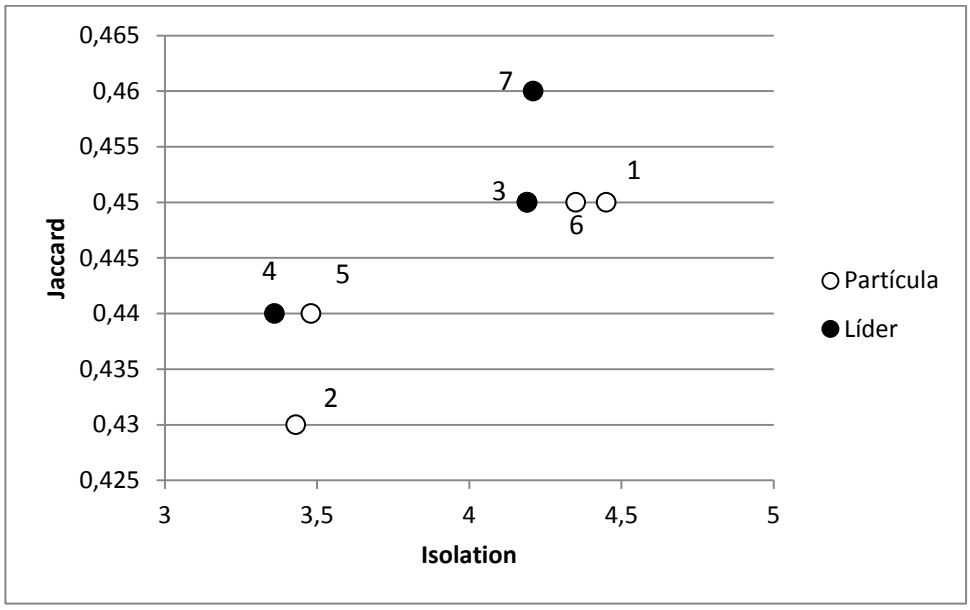

Figura 6.4: Soluções DLBCL-Tumor pós-MOPSO usando Classit, Isolation e Jaccard

6.3. Resultados da base Leukemia-ALL/AML

Nas bases Leukemia-ALL/AML, os resultados relevantes apareceram por parte dos métodos de seleção de atributos. Na Tabela 6.8 é mostrada a quantidade de atributos selecionados por seus respectivos métodos e observa-se que o C-FOCUS e o CFS selecionaram a mesma quantidade de genes.

Tabela 6.8: Quantidade de atributos selecionados para Leukemia-ALL/AML

Número do método	Quantidade de atributos
1	1
2	1
3	714
4	1782
5	3565
6	5347
7	7129

No Apêndice B, onde estão as listas dos genes mais selecionados para cada base de dados, revela que os dois método de seleção já citados selecionaram o mesmo atributo: o attribute6855. Por consequência, as soluções geradas a partir do C-FOCUS e do CFS serão as mesmas para qualquer método de agrupamento.

Além desses fatores, o mesmo atributo também foi selecionado pelas variações do *Relief*-F, mostrando-se de fato que é um atributo relevante na base de dados Leukemia-ALL/AML.

Esses resultados mostram também a importância de utilizar diversos métodos de seleção de atributos pertencentes a diferentes paradigmas. Assim, com diferentes conjuntos de atributos e observando os atributos em comum pode-se fazer uma avaliação com maior precisão. Além disso, como poder ser visto nos demais experimentos (Apêndice A), a redução de dimensionalidade realmente melhora os resultados e facilita na análise das bases de dados.

6.4. Resumo dos Resultados Experimentais

Nesta seção será apresentado um resumo sobre os resultados experimentais. Na Tabela 6.9 é mostrada, para cada base de dados de expressões gênicas, a quantidade de vezes que uma solução foi tida como ótima para um determinado método de otimização multiobjetivo.

Também é feita uma avaliação das informações obtidas e definida qual o melhor método de seleção de atributos para cada base. Essa avaliação é realizada com base na soma de todas as vezes que uma solução foi considerada ótima pelos métodos de otimização. Aquela que obteve a maior soma foi considerada a melhor para uma determinada base de dados.

Tabela 6.9: Resumo dos resultados experimentais

Bases de dados	Métodos de otimização multiobjetivo			Melhor método de seleção
	MOPSO	Lexicográfico	Formulação baseada em peso	
DLBCL-Stanford	Sol. 1 (6x) Sol. 2 (2x)	Sol. 1 (5x) Sol. 2 (1x)	Sol. 1 (4x) Sol. 2 (2x)	C-*FOCUS* (3 atributos)
DLBCL-Tumor	Sol. 1 (5x) Sol. 2 (3x) Sol. 3 (3x) Sol. 4 (3x) Sol. 5 (2x) Sol. 6 (2x) Sol. 7 (1x)	Sol. 1 (2x) Sol. 2 (2x) Sol. 4 (1x) Sol. 7 (1x)	Sol. 2 (2x) Sol. 3 (1x) Sol. 4 (1x) Sol. 6 (2x)	C-*FOCUS* (4 atributos) CFS (64 atributos)
DLBCL-Outcome	Sol. 1 (5x) Sol. 2 (2x) Sol. 3 (4x) Sol. 4 (1x) Sol. 5 (1x) Sol. 6 (2x) Sol. 7 (2x)	Sol. 1 (3x) Sol. 3 (3x)	Sol. 1 (3x) Sol. 2 (1x) Sol. 3 (1x) Sol. 7 (1x)	C-*FOCUS* (6 atributos)
DLBCL-NIH	Sol. 1 (6x) Sol. 2 (2x) Sol. 3 (1x) Sol. 5 (1x) Sol. 6 (1x) Sol. 7 (1x)	Sol. 1 (4x) Sol. 2 (1x) Sol. 3 (1x)	Sol. 1 (4x) Sol. 2 (2x)	C-*FOCUS* (5 atributos)
Leukemia-ALL/AML	Sol. 1 (5x) Sol. 2 (5x) Sol. 3 (4x) Sol. 4 (1x) Sol. 5 (2x) Sol. 6 (2x) Sol. 7 (1x)	Sol. 1 (4x) Sol. 2 (4x) Sol. 3 (1x) Sol. 7 (1x)	Sol. 1 (4x) Sol. 2 (4x) Sol. 3 (1x) Sol. 7 (1x)	C-*FOCUS* (1 atributo) CFS (1 atributo)
Leukemia MLL	Sol. 1 (4x) Sol. 2 (3x) Sol. 3 (5x) Sol. 4 (2x) Sol. 5 (4x) Sol. 6 (2x) Sol. 7 (1x)	Sol. 1 (2x) Sol. 2 (1x) Sol. 3 (1x) Sol. 5 (2x)	Sol. 2 (2x) Sol. 4 (2x) Sol. 5 (1x) Sol. 7 (1x)	*Relief*-F 50% (6291 atributos)

Como pode ser observada na Tabela 6.61, para a base DLBCL-Stanford sempre foi escolhida pelos métodos de otimização a solução 1 ou a solução 2, sendo que a solução 1 foi selecionada dez vezes mais que a solução 2.

Para a base DLBCL-Tumor, as soluções que representam o método C-*FOCUS* e CFS tiveram a mesma quantidade de vezes que foram consideradas ótimas pelos métodos de otimização.

A base DLBCL-Outcome, similar a base DLBCL-NIH, teve como melhor método de seleção o C-*FOCUS*. A diferença entre essas bases e a primeira base citada está na maior diversidade de soluções que foram selecionadas pelo MOPSO.

Com relação a base Leukemia-AML/ALL, a solução que apareceu mais vezes no conjunto das melhores soluções após a otimização multiobjetivo foi a de número 2, que representa o método CFS.

Diferente das demais, a base Leukemia-MLL teve o método *Relief*-F usando 50% da quantidade total de atributos da base como mais adequado.

.

Conclusões

Este trabalho objetivou avaliar, através da otimização multiobjetivo baseada em enxames, métodos de seleção de atributos aplicados em bases de dados de expressões gênicas obtidas pela técnica do microarranjo.

Primeiramente, é possível observar a partir dos experimentos que: a redução de dimensionalidade melhora os resultados das análises das bases de dados de expressões gênicas usando agrupamento. Essa afirmação é comprovada pela constante presença dos métodos de seleção de atributos entre as melhores soluções avaliadas pela otimização multiobjetivo, e pela ausência das soluções geradas pelas bases originais.

Entre os métodos de seleção, é destacada a grande quantidade de bases que selecionaram os métodos C-*FOCUS* e CFS como mais adequados. Esse fato reforça a afirmação de que poucos são os genes responsáveis pela ativação ou inibição de uma característica.

As soluções que representam os métodos de seleção são compostas por índices de validação de grupos. Os índices utilizados foram capazes de gerar, na maioria dos experimentos, os melhores valores para as bases que selecionaram as menores quantidades de atributos.

Nos experimentos que envolveram índices relativos, houve casos em que o MOPSO selecionou uma grande quantidade de soluções como ótimas, dificultando a escolha da melhor solução.

Para o índice Jaccard, por haver um conhecimento *a priori* sobre os agrupamentos originais, foram obtidos bons valores para grande quantidade de experimentos.

Esses índices foram obtidos pelo uso dos algoritmos K-*means*, ISODATA e Classit. O K-*means* obteve valores considerados bons para os melhores métodos de seleção, assim como o ISODATA, que em alguns casos chegou a agrupar da melhor forma possível as bases de dados.

Já o método hierárquico Classit, chegou a selecionar em alguns casos bases com grandes quantidades de atributos.

Também houveram casos em que o Classit gerou apenas um grande grupo. Com isso, os índices que fazem medição entre grupos obtiveram os piores valores possíveis.

Com relação ao método de otimização multiobjetivo baseado em enxame, o MOPSO, apesar de não ter acontecido situações de conflito entre índices de agrupamento, conseguiu selecionar as melhores soluções, quando comparados aos resultados obtidos pelos demais métodos de otimização utilizados (lexicográfico e baseado em fórmula de peso). Esses dois últimos métodos selecionaram soluções que estavam presentes no conjunto de líderes construído pela frente de Pareto, ao final da execução do MOPSO.

Foi possível observar também que o MOPSO, na maioria dos experimentos, resultou em um conjunto de líderes com mais de um componente. Esse fato mostra a importância de usar um tomador de decisões e selecionar a melhor solução de fato.

Como sugestão de futuros trabalhos, pode-se utilizar bases relacionadas a outras doenças, a fim de acumular conhecimento sobre o comportamento gênico quando elas ocorrem. Com isso, a proposta vai se tornar mais genérica, a ponto de servir como auxílio para cientistas e geneticistas. Por conta dessa generalização, será importante a utilização dos métodos de agrupamento pertencentes a outras abordagens, como o agrupamento baseado em densidade, para atender de forma eficiente a análises de expressões gênicas das doenças.

Pode-se também aplicar outras formas de reduzir a dimensionalidade das bases de dados, como os método de seleção *wrapper*, a fim de obter diferentes subconjuntos de atributos, já que foi mostrada que essa redução facilita e melhora a análise dessas bases.

Referências Bibliográficas

[Alizadeh et al., 2000] ALIZADEH, A. et al. *Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling*. Nature 4051, February, 2000, p.503-511.

[Arai & Bu, 2007] ARAI, K.; BU, X. *ISODATA Clustering With Parameter (Threshold for Merge and Split) Estimation Based on GA: Genetic Algorithm*. Science, Vol. 36, No. 1, 2007, p.17-23.

[Arauzo, Benitez & Castro, 2003] ARAUZO, A.; BENITEZ, J.M.; CASTRO, J.L. *C-FOCUS: A Continuos Extension of Focus.* Em: Advances in Soft Computing - Engineering, Design and Manufacturing, Springer, Vol. 35, No. 99, 2003, p.225-232.

[Armstrong et al., 2002] ARMSTRONG, S.A. et al. *MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia*. Nature Genetics, No. 30, 2002, p.41-47.

[Au et al., 2005] AU, W.H.; CHAN, K.C.C.; WONG, A.K.C.; WANG, Y. *Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 2, 2005, p.83-101.

[Azuaje & Bolshakova, 2002] AZUAJE, F.; BOLSHAKOVA, N. *Clustering Genome Expression Data: Design and Evaluation Principles, Understanding and Using Microarray Analysis Techniques: A Practical Guide Springer*. Berrar, D.; Dubitzky, W.; Granzow, M. (eds.), Verlag, 2002.

[Bolshakova & Azuaje, 2003] BOLSHAKOVA, N.; AZUAJE, F. *Cluster Validation Techniques for Genome Expression Data Classification. Signal Processing*. Vol. 83, No. 4, 2003, p.825-833.

[Bolshakova & Azuaje, 2006] BOLSHAKOVA, N.; AZUAJE, F. *Estimating the Number of Clusters in DNA Microarray Data*. Methods Inf. Med, Vol. 45, No. 2, pp. 153--157, 2006.

[Borges, 2006] BORGES, H.B. *Redução de Dimensionalidade em Bases de Dados de Expressão Gênica*. Dissertação de Mestrado, PPGIa - PUCPR, 2006.

[Brown & Botstein, 1999] BROWN, P.; BOTSTEIN, D. *Exploring The New World of The Genome With DNA Microarrays.* Nature Genetics, Vol. 21, 1999, p.33-37.

[Carvalho & Pozo, 2009] CARVALHO, A.B.; POZO, A.T.R. *Otimização por Nuvem de Partículas Multiobjetivo na Aprendizagem Indutiva de Regras: Extensões e Aplicações*. Dissertação de mestrado, Universidade Federal do Paraná, 2009.

[Chuang et al., 2008] CHUANG, L.Y.; CHANG, H.W.; TU, C.J.; YANG, C.H. *Computational Biology and Chemistry*, Elsevier, Vol. 32, 2008, p.29-38.

[Coello, Lamont & Vldhuizen, 2007] COELLO, C.A.; LAMONT, G.B.; VELDHUIZEN, D.A.V. *Evolutionary Algorithms for Solving Multi-Objective Problems*, Segunda Edição, Springer-Verlag, 2007.

[Dash et al., 2002] DASH, M.; CHOI, K.; SCHEUERMANN, P.; LIU, H. *Feature Selection for Clustering – A Filter Solution*. Second International Conference on Data Mining, 2002, p.115-122.

[Devaney & Ram, 1997] DEVANEY, M.; RAM, A. *Efficient Feature Selection in Conceptual Clustering.* 14 th International Conference on Machine Learning, 1997, p.92-97.

[Dy, 2008] DY, J. *Unsupervised Feature Selection.* Em: Computational Methods of Feature Selection, Liu, H. & Motoda, H. (eds.), Chapman & Hall/CRC, 2008, p.19-39.

[D'Haeseleer, 2005] D'HAESELEER, P. *How Does Gene Expression Clustering Work?* Nature Biotechnology, Vol. 23, No. 12, 2005, p.1499-1501.

[Eberhart & Shi, 2001] EBERHART, R.C.; SHI, Y. *Particle Swarm Optimization: Developments, Applications and resource.* IEEE International Conference Evolutionary Computation, vol. 1, 2001, p.81-86.

[Faceli, Carvalho & Souto, 2005] FACELI, K.; CARVALHO, A.; SOUTO, M. *Validação de Algoritmos de Agrupamento*. Relatórios Técnicos do ICMC, 2005.

[Freitas, 2004] FREITAS, A.A. *A critical review of Multi-Objective Optimization in Data Mining: A Position Paper*. Universidade de Kent, UK, 2004.

[Garcia & Nievola, 2011] GARCIA, R.; NIEVOLA, J.C. Uso de Critérios Multiobjetivo Baseados em Enxames na Escolha dos Melhores Métodos para Seleção de Atributos em Microarranjos Gênicos. VIII Encontro Nacional de Inteligência Artificial (ENIA), Natal, 2011.

[Garcia, Paraiso & Nievola, 2011] GARCIA, R.; PARAISO, E.C.; NIEVOLA, J.C. *Multiobjective Optimization of Indexes Obtained by Clustering for Feature Selection Methods Evaluation in Genes Expression Microarrays*. 12th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Norwich, UK, 2011.

[Golub et al., 1999] GOLUB, T. et al. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Science, 1999, p.531-537.

[Hall, 2000] HALL, M. *Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning*. 17th International Conference on Machine Learning, 2000, p.359-366.

[Jain & Dubes, 1988] JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Prentice-Hall. 1988.

[Jain, Murty & Flynn, 1999] JAIN., A. K.; MURTY, M. N.; FLYNN, P. J. *Data Clustering: A Review*. ACM Computing Surveys, Vol. 31, No. 3, 1999, p.264-323.

[Kaufmann & Michalski, 1999] KAUFMANN, K.A.; MICHALSKI, R.S. *Learning from inconsistent and noisy data: the AQ18 approach*. Foundations of Intelligent Systems, Springer, 1999, p.411-419.

[Kennedy & Eberhart, 1995] KENNEDY, J.; EBERHART, R.C. *Particle Swarm Optimization*. IEEE International Conference on Neural Networks, IEEE Press, 1995, p.1942-1948.

[Kohavi & John, 1997] KOHAVI, R.; JOHN, G.H. *Wrappers for Feature Subset Selection*. Artificial Intelligence, 1997, p.273-324.

[Kononenko & Sikonja, 2008] KONONENKO, I.; SIKONJA, M.R. *Non-Myopic Feature Quality Evaluation with (R)ReliefF*. Em: Computational Methods of Feature Selection, Liu, H. & Motoda, H (eds.), Chapman & Hall/CRC, 2008, p.169-191.

[Liu & Yu, 2005] LIU, H.; YU, L. *Toward Integrating Feature Selection Algorithms for Classification and Clustering*. IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, 2005, p.491-502.

[Maulik & Bandyopadhyay, 2002] MAULIK, U.; BANDYOPADHYAY, S. *Performance Evaluation of Some Clustering Algorithms and Validity Indices*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, 2002, p.1650-1654.

[Mostaghim & Teich, 2003] MOSTAGHIM, S.; TEICH, J. *Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO)*. 2003 IEEE Swarm Intelligence Symposium, 2003, p.26-33.

[NIH, 2001] NIH - National Institutes of Health. *Genetic Basic*. Disponível em: http://www.nigms.nih.gov. 2001.

[Pappa et al, 2004] PAPPA, G.L.; FREITAS, A.A.; KAESTNER, C.A.A. *Multi-Objective Algorithms for Attribute Selection in Data Mining*. Applications of Multi-Objective Evolutionary Algorithms, Coello, C.A.; Lamont, G.B. (ed.), World Scientific, 2004, p.603-626.

[Quackenbush, 2001] QUACKENBUSH, J. *Computational Analysis of Microarray Data*. Nature Reviews – Genetics, Vol. 2, 2001, p.418-427.

[Rosenwald et al., 2002] ROSENWALD, A. et al. *The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma*. N Engl J Med, Vol. 346, No. 25, 2002, p.1937-1947.

[Saeys, Inza & Larrañaga, 2007] SAEYS, Y.; INZA, I.; LARRAÑAGA, P. *A Review of Feature Selection Techniques in Bioinformatics*. Bioinformatics Advance Access, Vol. 23, No. 19, 2007, p.2507-2517.

[Shipp et al., 2002] SHIPP, M. et al. *Diffuse large B-cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning*. Nature Medicine, Vol. 8, No. 1, 2002, p.68-74.

[Sierra & Coello, 2006] SIERRA, M.R.; COELLO, C.A.C. *Multi-objective Particle Swarm Optimizers: A Survey of The State-of-the-art*. International Journal of computational Intelligence Research, Vol. 2, 2006, p.287-308.

[Sunaga, 2006] SUNAGA, D.Y. *Aplicação de Técnicas de Validação Estatística e Biológica em Agrupamento de Dados de Expressão Gênica*. Dissertação de Mestrado, PPGIa - PUCPR, 2006.

[Suresh et al, 2009] SURESH, K.; KUNDU, D.; GHOSH, S.; DAS, S. *Data Clustering Using Multi-objective Differential Evolution Algorithm*. Fundamenta Informaticae. IOS Press, Vol. 21, 2009, p.1001-1024.

[Tan, Steinbach & Kumar, 2006] TAN, P.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. Addison-Wesley, 2006.

[Theodoridis & Koutroumbas, 2003] THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*, Segunda Edição, Elsevier, 2003, p.531-533.

[Xing, Jordan & Karp, 2001] XING, E.; JORDAN, M.; KARP, R. *Feature Selection for High-Dimensional Genomic Microarray Data*. 18th International Conference on Machine Learning, 2001, p.601-608.

[Xu & Wunsch, 2005] XU, R.; WUNSCH, D. *Survey of Clustering Algorithms*. IEEE Transactions on Neural Networks, Vol. 16, No. 3, 2005, p.645-678.

[Yu & Liu, 2004] YU, L.; LIU, H. *Redundancy Based Feature Selection for Microarray Data*. Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2004, p.737-742.

[Zhao & Karypis, 2003] ZHAO, Y.; KARYPIS, G. *Clustering in Life Sciences*. Em: Functional Genomics: Methods and Protocols, Khodursky, A. & Brownstein, M. (eds.), 2003, p.183-218.

Apêndice A

Resultados Experimentais

Nesta extensão do trabalho serão vistos todos os resultados que foram obtidos pelos experimentos explicados no capítulo 6. As Tabelas A.1, A.11, A.21, A.31, A.41 e A.51 mostram a quantidade de genes que foram selecionados pelos métodos de seleção de atributos.

Os resultados serão apresentados em tabelas, contendo a identificação do método de seleção e os valores dos objetivos nos momentos antes e pós realização do MOPSO. Comentários sobre as soluções escolhidas pelos métodos lexicográfico e baseado em fórmula de peso serão feitos para comparação com os líderes escolhidos pelo MOPSO e auxiliar na escolha da melhor solução para cada base de dados.

Na otimização lexicográfica, a sequência de preferência de índices, quando estes estiverem presentes, obedecerá a sequência já apresentada no capítulo anterior.

Além das tabelas, os gráficos ilustrando as posições das soluções também obedecem a seguinte numeração para identificar a qual método de seleção elas pertencem:

1- Solução obtida pelo método C-FOCUS;
2- Solução obtida pelo método CFS;
3- Solução obtida pelo método *Relief*-F com 10% do tamanho original;
4- Solução obtida pelo método *Relief*-F com 25% do tamanho original;
5- Solução obtida pelo método *Relief*-F com 50% do tamanho original;
6- Solução obtida pelo método *Relief*-F com 75% do tamanho original;
7- Solução obtida pela base original;

A.1. Resultados da Base DLBCL-Stanford

Tabela A.1: Quantidade de atributos selecionados para DLBCL-Stanford

Número do método	Quantidade de atributos
1	3
2	47
3	402
4	1006
5	2013
6	3020
7	4026

A.1.1. K-means

Tabela A.2: Soluções DLBCL-Stanford usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**1.52**	0.52
2	4.34	**0.77**
3	12.34	0.64
4	17.93	0.54
5	21.18	0.32
6	24.62	0.26
7	27.4	0.26

Na Tabela A.2 estão as soluções das bases DLBCL-Stanford ao serem avaliadas pelos critérios *Isolation* e Jaccard com o método K-means de agrupamento. Observa-se que as menores bases resultaram nos melhores valores, sendo que a solução 1 otimizou o índice *Isolation* e a solução 2 otimizou o índice Jaccard.

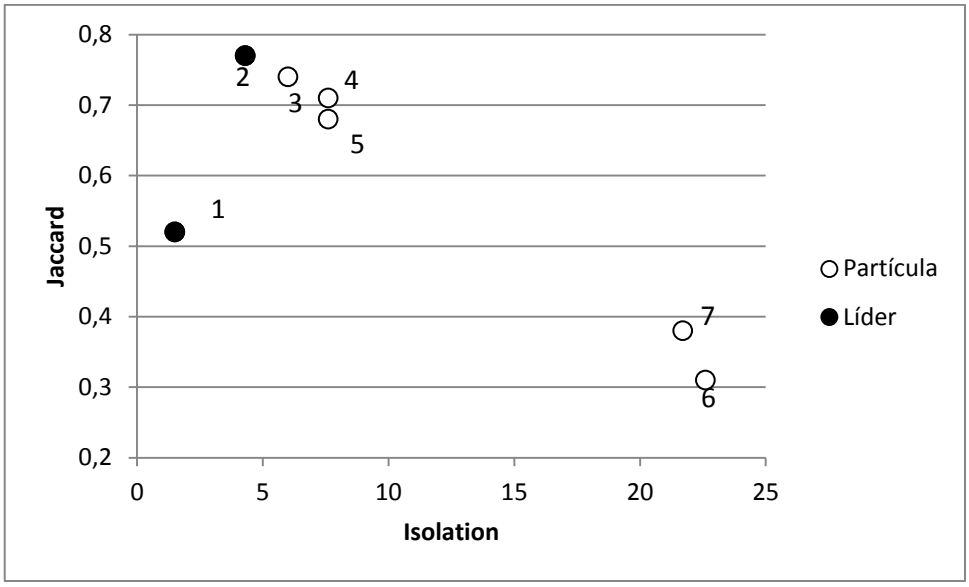

Figura A.1: Soluções DLBCL-Stanford pós-MOPSO usando K-means, Isolation e Jaccard

Os líderes escolhidos pela frente de Pareto ao fim do MOPSO, como podem ser vistos na Figura A.1, foram as soluções 1 e 2. A solução 1 domina as soluções 6 e 7, assim como a solução 2, que também domina as soluções 3, 4 e 5. Para os métodos lexicográfico e baseado em fórmula de peso, o melhor método de seleção de atributos foi o CFS, representada pela solução 2.

Tabela A.3: Soluções DLBCL-Stanford usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	**1.46**	**1.91**	**0.44**
2	**0.41**	1.63	3.54	0.36
3	0.39	1.97	5.0	0.29
4	0.4	2.61	5.76	0.2
5	0.38	3.0	6.0	0.15
6	0.36	2.99	6.31	0.12
7	0.37	3.14	6.15	0.13

Na Tabela A.3 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 2 otimiza o índice Dunn e a solução 1 otimiza os índices Davies-Bouldin, C e Silhueta.

Tabela A.4: Soluções DLBCL-Stanford pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	**1.46**	**1.91**	**0.44**
2	**0.41**	1.63	3.54	0.36
3	0.4	1.89	4.69	0.3
4	0.4	2.38	5.25	0.23
5	0.4	2.13	4.46	0.28
6	0.37	2.77	5.87	0.16
7	0.38	2.48	5.02	0.23

A Tabela A.4 mostra para onde o MOPSO deslocou as soluções. Pelo fato das bases menores terem uma diferença considerável no número de atributos selecionados, os índices de agrupamento tiveram uma grande melhora, não permitindo que a otimização aproximasse as demais soluções das melhores. Por isso, os últimos líderes escolhidos pela frente de Pareto foram as soluções 1 e 2.

A.1.2. ISODATA

Tabela A.5: Soluções DLBCL-Stanford usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**1.87**	**1.0**
2	2.95	0.36
3	18.57	0.31
4	27.31	0.32
5	25.54	0.33
6	45.13	0.30
7	52.87	0.30

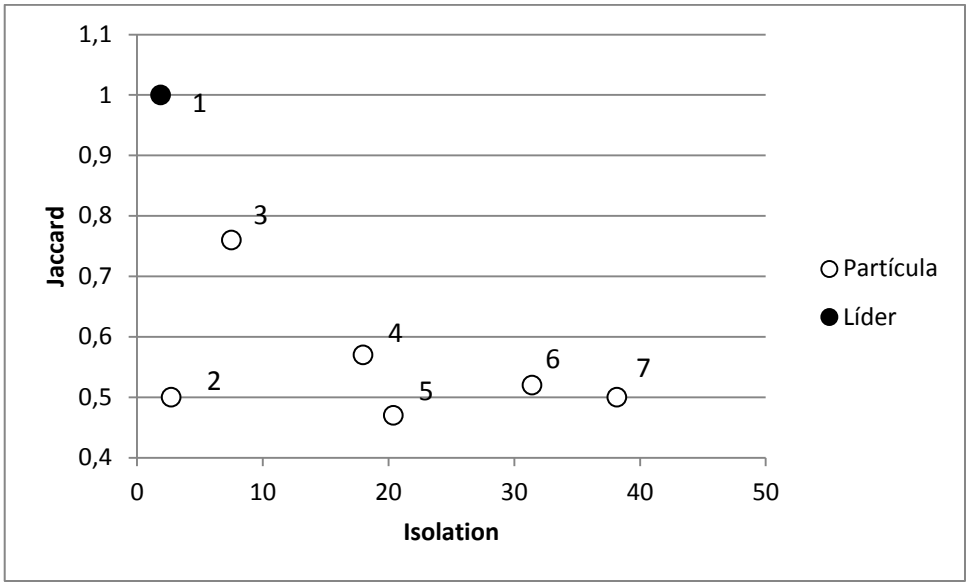

Figura A.2: Soluções DLBCL-Stanford pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Tabela A.5 estão as soluções das bases DLBCL-Stanford ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento ISODATA. A solução 1 foi estritamente melhor que as demais soluções, ao otimizar ambos os índices. Esta solução, por sua vez, continuou sendo a única escolhida pelo MOPSO como ótima (Figura A.2). O mesmo ocorreu nos métodos de otimização multiobjetivo lexicográfico e baseado em fórmula de peso.

Tabela A.6: Soluções DLBCL-Stanford usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.64**	**1.22**	**1.55**	**0.26**
2	0.27	1.80	3.51	0.04
3	0.32	1.66	5.12	0.08
4	0.26	1.35	5.76	0.02
5	0.36	1.64	6.0	0.01
6	0.28	1.62	6.44	0.07
7	0.29	1.38	5.68	0.01

Na Tabela A.6 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 1 otimiza todos os índices relativos.

Tabela A.7: Soluções DLBCL-Stanford pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.64**	**1.22**	**1.55**	**0.26**
2	0.37	1.65	3.01	0.09
3	0.34	1.63	4.93	0.02
4	0.53	1.26	2.80	0.19
5	0.62	1.25	1.88	0.24
6	0.57	1.30	2.56	0.20
7	0.38	1.33	4.57	0.08

Após o processamento do MOPSO, Tabela A.7, não houve alteração da melhor solução, sendo selecionada a solução 1. A mesma solução foi selecionada pelos métodos lexicográfico e baseado na fórmula de peso.

6.1.3. Classit

Tabela A.8: Soluções DLBCL-Stanford usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.0**	**0.48**
2	85.26	0.30
3	322.48	0.20
4	357.68	0.21
5	466.6	0.21
6	345.05	0.20
7	1396.55	0.14

Na Tabela A.8 estão as soluções das bases DLBCL-Stanford ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento hierárquico Classit. A melhor solução foi a de número 1. Seu valor 0.0 para o índice *Isolation* é explicada pelo fato de ter sido construído apenas um grupo, como se todas as instâncias da base de dados tivessem

comportamentos similares e pertencessem a pessoas com as mesmas características. Nota-se também o baixo valor do índice Jaccard para essa solução.

Consequentemente, após a execução do MOPSO (Figura A.3), assim como nos outros métodos de otimização multiobjetivo, a única solução selecionada como ótima foi a obtida pelo método de seleção C-*FOCUS*. Por outro lado, a solução que representa a base original foi classificada como a pior solução, bem distante das demais.

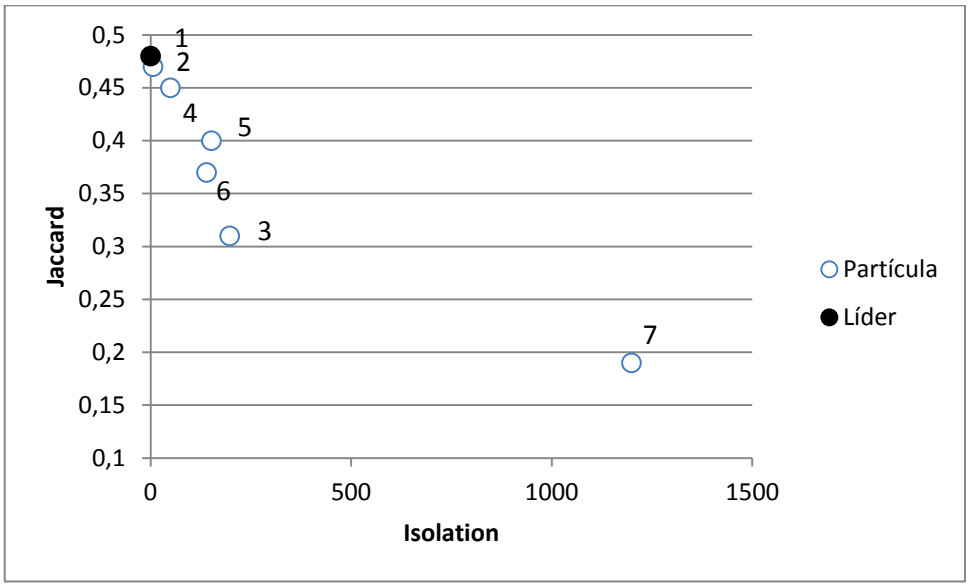

Figura A.3: Soluções DLBCL-Stanford pós-MOPSO usando Classit, Isolation e Jaccard

Tabela A.9: Soluções DLBCL-Stanford usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.79**	**0.0**	**-3.22**	**1.0**
2	0.35	1.67	1.77	-0.47
3	0.41	1.80	2.28	-0.28
4	0.40	1.69	2.75	-0.22
5	0.34	1.73	3.64	-0.52
6	0.38	1.73	3.50	-0.38
7	0.37	1.62	2.90	-0.49

Na Tabela A.9 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 1 otimiza todos os índices relativos. Por não se tratar de um agrupamento hiper-esférico, os índices Silhueta e Davies-Bouldin podem ser considerados não realistas.

Assim como no experimento anterior, envolvendo os índices *Isolation* e Jaccard, todos os métodos de otimização selecionaram a solução 1 como a melhor, como pode ser visto na Tabela A.10.

Tabela A.10: Soluções DLBCL-Stanford pós-MOPSO usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.79**	**0.0**	**-3.22**	**1.0**
2	1.61	0.16	-2.9	0.85
3	1.47	0.32	-2.64	0.76
4	1.13	0.62	-2.04	0.55
5	1.68	0.10	-3.02	0.90
6	1.08	0.69	-1.94	0.44
7	1.61	0.16	-2.90	0.85

A.2. Resultados da Base DLBCL-Tumor

Tabela A.11: Quantidade de atributos selecionados para DLBCL-Tumor

Número do método	Quantidade de atributos
1	4
2	64
3	713
4	1782
5	3565
6	5347
7	7129

A.2.1. K-means

Tabela A.12: Soluções DLBCL-Tumor usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.43**	0.48
2	1.28	**0.78**
3	4.70	0.42
4	6.77	0.39
5	8.62	0.40
6	6.78	0.45
7	7.42	0.44

Na Tabela A.12 estão as soluções das bases DLBCL-Tumor ao serem avaliadas pelos critérios *Isolation* e Jaccard. Observa-se que as menores bases resultaram nos melhores valores, sendo que a solução 1 otimizou o índice *Isolation* e a solução 2 otimizou o índice Jaccard.

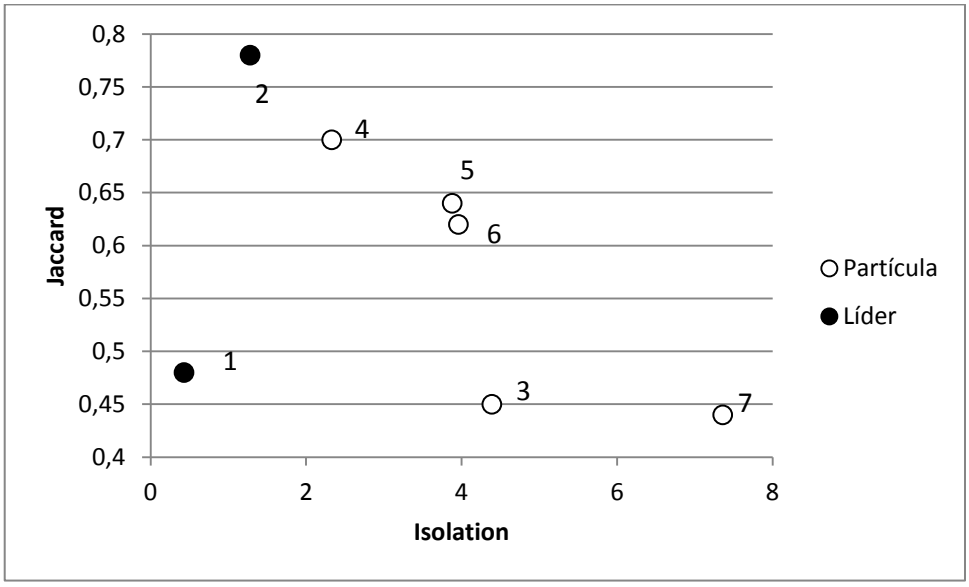

Figura A.4: Soluções DLBCL-Tumor pós-MOPSO usando K-means, Isolation e Jaccard

Os líderes escolhidos pela frente de Pareto ao fim do MOPSO, como podem ser vistos na Figura A.4, foram as soluções 1 e 2. A solução 1 domina as soluções 3 e 7, assim como a

solução 2, que também domina as soluções 4, 5 e 6. Para os métodos lexicográfico e baseado em fórmula de peso, o melhor método de seleção de atributos foi o CFS, representada pela solução 2.

Na Tabela A.13 é possível observar as soluções obtidas pelos critérios relativos, em que as soluções 4 e 5 otimizam o índice Dunn, a solução 1 otimiza os índices Davies-Bouldin e C, e a solução 4 otimiza também o índice Silhueta.

Tabela A.13: Soluções DLBCL-Tumor usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.09	**1.28**	**3.49**	0.15
2	0.24	1.97	4.87	0.20
3	0.30	2.01	4.54	0.22
4	**0.31**	2.09	4.23	**0.25**
5	**0.31**	2.23	4.14	0.17
6	0.20	3.35	5.08	0.05
7	0.19	3.46	4.99	0.05

Tabela A.14: Soluções DLBCL-Tumor pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.09	**1.28**	**3.49**	0.15
2	0.24	1.97	4.87	0.20
3	0.30	2.01	4.54	0.22
4	**0.31**	2.09	4.23	**0.25**
5	**0.31**	2.23	4.14	0.17
6	**0.31**	2.28	4.18	0.16
7	0.22	3.14	4.77	0.08

A Tabela A.14 mostra como o MOPSO desloca as soluções. As únicas soluções não tidas com ótimas foram a de número 6, por ser dominada pelas soluções 4 e 5, e a solução 7,

que é dominada pelas soluções 3, 4 e 5. Os métodos lexicográfico e baseado em fórmula de peso selecionaram a solução do método *Relief*-F 25% como a melhor.

A.2.2. ISODATA

Tabela A.15: Soluções DLBCL-Tumor usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.23**	0.39
2	1.15	**0.54**
3	4.51	0.41
4	7.20	0.40
5	8.50	0.39
6	9.43	0.39
7	10.69	0.46

Na Tabela A.15 estão as soluções das bases DLBCL-Tumor ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento ISODATA. A solução 1 conseguiu otimizar o índice *Isolation* e a solução 2 otimizou o índice Jaccard.

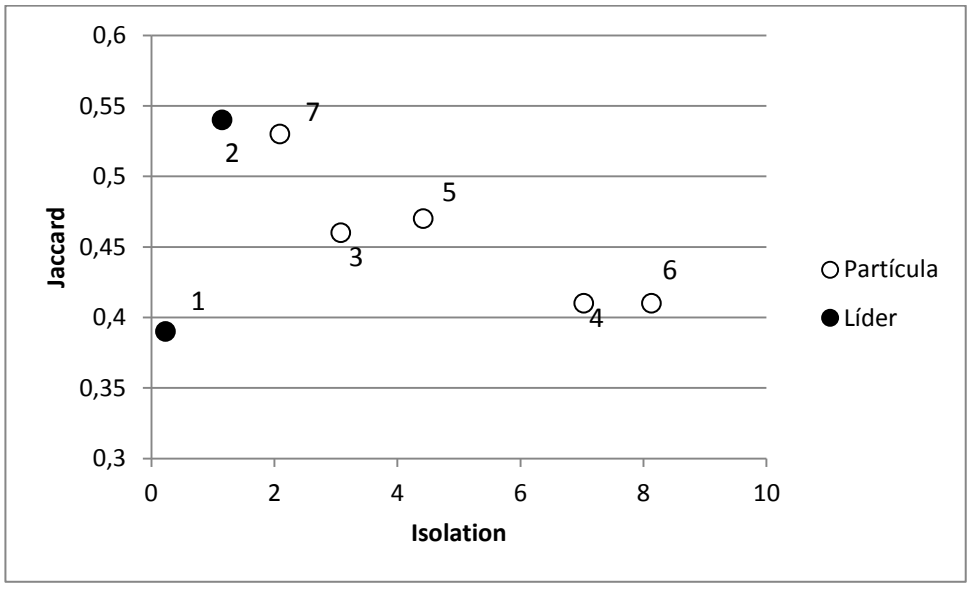

Figura A.5: Soluções DLBCL-Tumor pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Figura A.5, a solução 2 domina todas as outras, exceto a solução 1, que possui menor valor do índice Isolation. Nos métodos de otimização multiobjetivo lexicográfico e baseado em fórmula de peso foi selecionada a solução 2.

Tabela A.16: Soluções DLBCL-Tumor usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.04	4.23	**1.01**	0.15
2	0.21	2.54	4.28	0.16
3	**0.33**	**1.95**	5.21	0.15
4	0.32	1.99	4.93	0.15
5	0.31	2.19	4.34	0.15
6	0.27	2.35	3.70	**0.22**
7	0.23	2.36	4.28	0.12

Na Tabela A.16 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 3 otimiza os índices Dunn e Davies-Bouldin, a solução 1 otimiza o índice C e a solução 6 otimiza o índice Silhueta.

Como a solução 1 obteve o melhor valor e além da tolerância dos outros valores para o índice C, foi escolhida como melhor solução pelo método lexicográfico. Para o método baseado na fórmula de peso, a melhor solução foi a de número 6.

Tabela A.17: Soluções DLBCL-Tumor pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.04	4.23	**1.01**	0.15
2	0.23	2.48	4.11	0.18
3	**0.33**	**1.95**	5.21	0.15
4	0.32	1.99	4.93	0.15
5	0.31	2.19	4.34	0.15
6	0.27	2.35	3.70	**0.22**
7	0.31	2.20	4.34	0.15

Após o processamento do MOPSO, a solução 2 foi dominada pela solução 6 e a solução da base original foi dominada pela solução 5, por ter menor valor do índice Davies-Bouldin. As soluções 2 e 7 foram as únicas a não serem classificadas como ótimas Tabela A.17.

A.2.3. Classit

Tabela A.18: Soluções DLBCL-Tumor usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	4.42	0.45
2	3.50	0.43
3	4.14	0.42
4	**3.36**	0.44
5	3.46	0.44
6	4.59	0.44
7	4.21	**0.46**

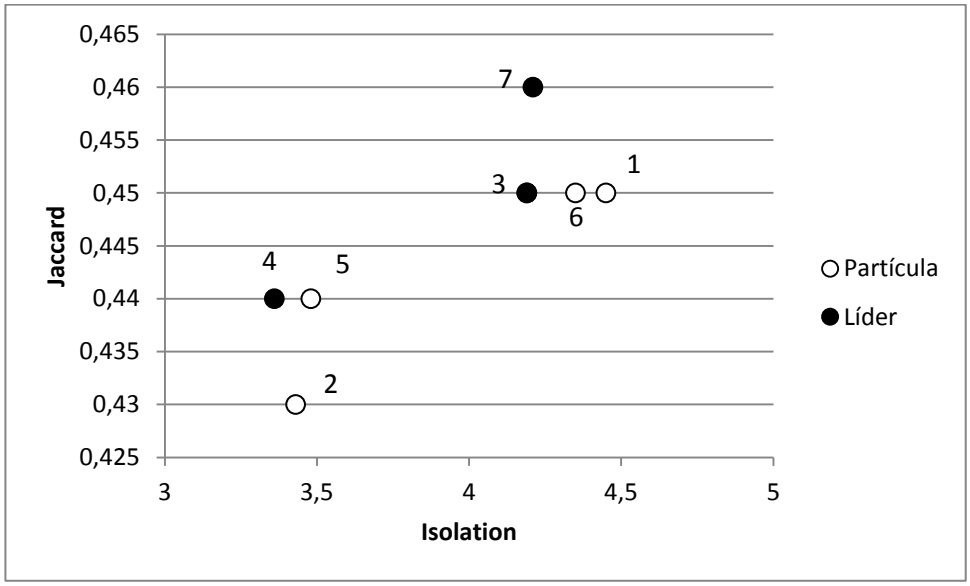

Figura A.6: Soluções DLBCL-Tumor pós-MOPSO usando Classit, Isolation e Jaccard

Na Tabela A.18 estão as soluções das bases DLBCL-Tumor ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento Classit. A solução 7 conseguiu otimizar o índice Jaccard e a solução 4 otimizou o índice *Isolation*.

Na Figura A.6, a solução 4 domina as soluções 4 e 5, e as soluções 3 e 7 dominam as soluções 1 e 6. O método lexicográfico escolheu a solução 7, que representa a base original, e o método baseado em fórmula de peso escolheu a solução 3.

Tabela A.19: Soluções DLBCL-Tumor usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.196**	6.08	**4.99**	0.012
2	0.191	7.78	**4.99**	0.008
3	0.191	6.71	**4.99**	0.012
4	0.193	8.21	**4.99**	0.008
5	0.191	7.91	**4.99**	0.010
6	0.194	**5.98**	**4.99**	**0.013**
7	0.0	6.36	5.17	0.009

Na Tabela A.19 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 1 otimiza o índice Dunn e a solução 6 otimiza os índices Davies-Bouldin e Silhueta. Para o índice C, todas as soluções, exceto a solução 7, foram capazes de otimizá-la.

Tabela A.20: Soluções DLBCL-Tumor pós-MOPSO usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.196**	6.08	**4.99**	0.012
2	0.194	6.94	**4.99**	0.010
3	0.193	6.46	**4.99**	0.012
4	0.194	7.38	**4.99**	0.010
5	0.192	7.63	**4.99**	0.010
6	0.194	**5.98**	**4.99**	**0.013**
7	0.043	6.30	5.13	0.009

Após o processamento do MOPSO, Tabela A.20, a solução 1 otimiza o índice Dunn e a solução 6 otimiza os índices Silhueta e Davies-Bouldin. Por isso, essas duas soluções foram eleitas líderes pelo MOPSO. O método baseado na fórmula de peso selecionou a solução 6, enquanto que o método lexicográfico selecionou a solução 1.

A.3. Resultados da Base DLBCL-Outcome

Tabela A.21: Quantidade de atributos selecionados para DLBCL-Outcome

Número do método	Quantidade de atributos
1	6
2	35
3	713
4	1782
5	3565
6	5347
7	7129

A.3.1. K-means

Tabela A.22: Soluções DLBCL-Outcome usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.45**	0.40
2	0.89	0.35
3	4.81	**0.41**
4	7.16	0.39
5	10.19	0.40
6	11.31	0.38
7	11.23	0.32

Na Tabela A.22 estão as soluções das bases DLBCL-Outcome ao serem avaliadas pelos critérios *Isolation* e Jaccard. Observa-se que a solução 1 otimizou o índice *Isolation* e a solução 3 otimizou o índice Jaccard.

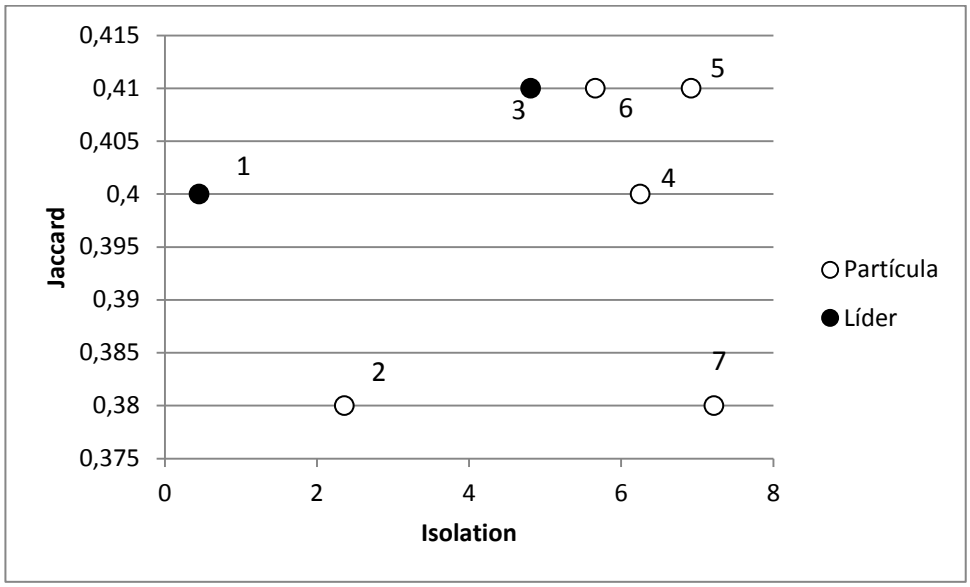

Figura A.7: Soluções DLBCL-Outcome pós-MOPSO usando K-means, Isolation e Jaccard

Os líderes escolhidos pela frente de Pareto ao fim do MOPSO, como podem ser vistos na Figura A.7, foram as soluções 1 e 3. A solução 1 domina as soluções 2 e 7, assim como a solução 3, que também domina as soluções 4, 5 e 6. O método lexicográfico selecionou a solução 3, enquanto que o método baseado em fórmula de peso selecionou a solução 1.

Tabela A.23: Soluções DLBCL-Outcome usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.12	**1.65**	**2.33**	**0.36**
2	0.35	2.49	4.70	0.09
3	**0.42**	2.28	8.78	0.04
4	0.39	2.30	6.85	0.06
5	0.40	2.26	7.02	0.06
6	0.27	2.44	6.15	0.06
7	0.36	2.64	4.41	0.18

Na Tabela A.23 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 3 otimizou o índice Dunn, a solução 1 otimizou os índices Davies-Bouldin, C e Silhueta.

Tabela A.24: Soluções DLBCL-Outcome pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.12	**1.65**	**2.33**	**0.36**
2	0.35	2.49	4.70	0.09
3	**0.42**	2.28	8.78	0.04
4	0.39	2.30	6.85	0.06
5	0.40	2.26	7.02	0.06
6	0.27	2.44	6.15	0.06
7	0.36	2.64	4.41	0.18

A Tabela A.24 mostra como o MOPSO desloca as soluções. Pelo fato de a solução 1 obter o pior valor para o índice Dunn, apesar de otimizar os demais índices, ela não domina nenhuma solução. Além disso, nenhuma solução foi dominada e, assim, todas foram eleitas líderes e não foram deslocadas. Os métodos lexicográfico e baseado em fórmula de peso selecionaram a solução do método C-*FOCUS* como a melhor.

A.3.2. ISODATA

Tabela A.25: Soluções DLBCL-Outcome usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.37**	**0.37**
2	0.67	0.34
3	3.59	**0.37**
4	5.88	0.33
5	8.13	0.32
6	9.67	0.32
7	11.27	0.32

Na Tabela A.25 estão as soluções das bases DLBCL-Outcome ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento ISODATA. A solução 1 conseguiu otimizar os índices *Isolation* e Jaccard, esse último em conjunto com a solução 3.

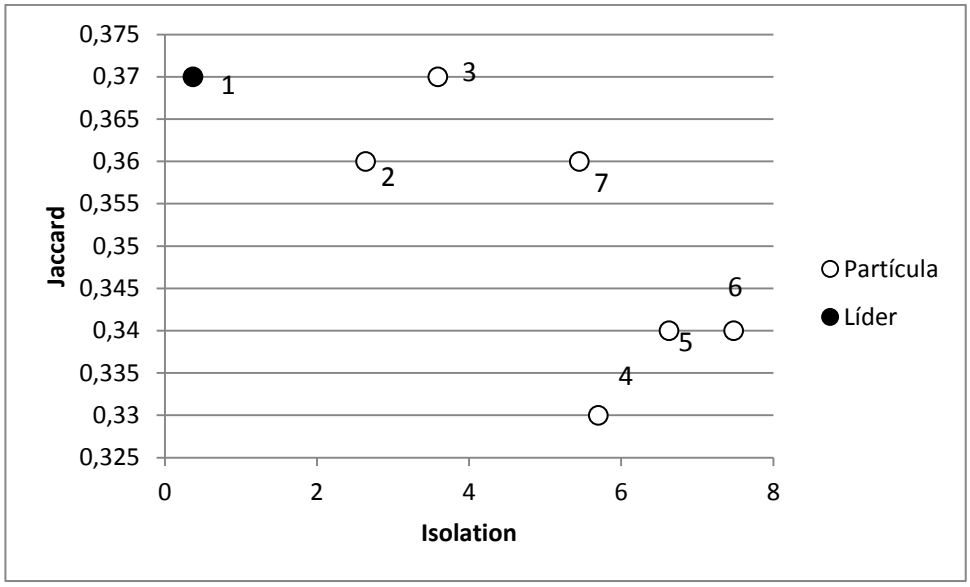

Figura A.8: Soluções DLBCL-Outcome pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Figura A.8, a solução 1 domina todas as outras, se tornando o único líder do enxame. O método otimização multiobjetivo baseado em fórmula de peso também selecionou a solução 1, enquanto que o método lexicográfico selecionou a solução 3.

Tabela A.26: Soluções DLBCL-Outcome usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	3.43	**1.80**	0.22
2	**0.38**	1.86	3.84	0.32
3	0.36	**0.99**	3.14	**3.59**
4	0.34	2.56	4.38	0.20
5	0.34	2.56	4.23	0.19
6	0.33	2.54	4.39	0.17
7	0.37	2.66	4.68	0.14

Na Tabela A.26 é possível observar que a solução 2 otimiza o índice Dunn, a solução 3 otimiza os índices Davies-Bouldin e Silhueta, e a solução 1 otimiza o índice C.

Pelo fato de a solução 1 ter obtido o melhor valor e além da tolerância dos outros valores para o índice C, foi escolhida como melhor solução pelo método lexicográfico. Para o método baseado na fórmula de peso, a melhor solução foi a de número 2.

Tabela A.27: Soluções DLBCL-Outcome pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.08	3.43	**1.80**	0.22
2	**0.38**	1.86	3.84	0.32
3	0.36	**0.99**	3.14	**3.59**
4	0.35	2.35	4.22	0.23
5	0.35	2.39	4.14	0.22
6	0.34	2.37	4.25	0.20
7	0.37	2.32	4.33	0.22

Após o processamento do MOPSO (Tabela A.27), as soluções 4 e 5 foram dominadas pelas soluções 2 e 3. A solução 1, por otimizar o índice C, não é dominado por ninguém, sendo considerada solução ótima juntamente com as soluções 2 e 3.

A.3.3. Classit

Tabela A.28: Soluções DLBCL-Outcome usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**3.79**	0.34
2	3.88	0.32
3	5.49	0.32
4	4.09	0.32
5	4.12	0.33
6	3.60	**0.32**
7	3.23	0.32

Na Tabela A.28 estão as soluções das bases DLBCL-Outcome ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento Classit. A solução 1 conseguiu otimizar o índice Jaccard, e a solução 6 otimiza o índice *Isolation*.

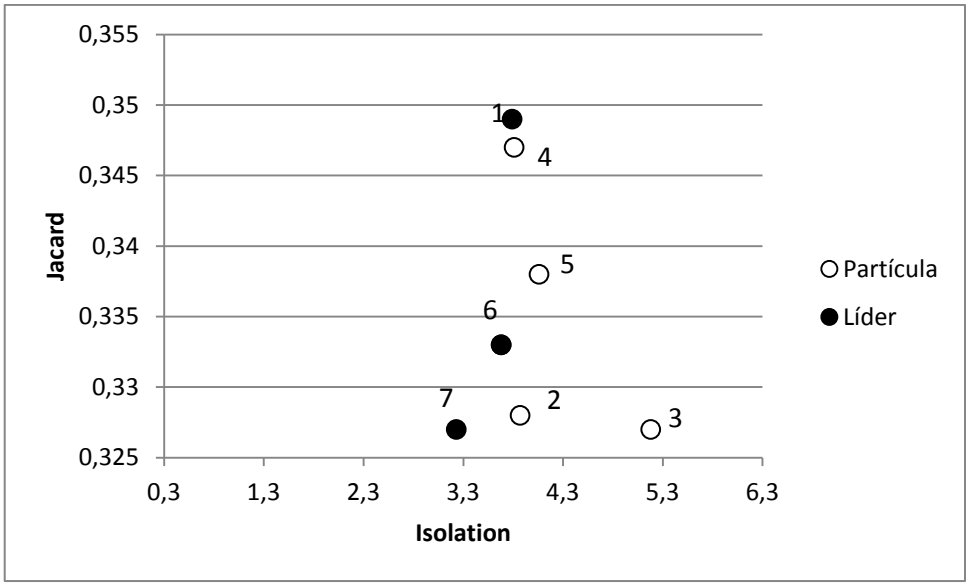

Figura A.9: Soluções DLBCL-Outcome pós-MOPSO usando Classit, Isolation e Jaccard

Na Figura A.9, a solução 1 domina todas as outras, exceto as soluções 6 e 7, que também formam o conjunto de líderes. O método otimização multiobjetivo baseado em fórmula de peso também selecionou a solução 7, enquanto que o método lexicográfico selecionou a solução 1.

Tabela A.29: Soluções DLBCL-Outcome usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.27	8.20	**4.43**	0.02
2	0.27	8.08	**4.43**	0.02
3	**0.28**	**5.60**	**4.43**	**0.05**
4	0.27	7.61	**4.43**	0.02
5	0.25	7.52	**4.43**	0.03
6	0.25	8.65	**4.43**	0.02
7	0.26	9.70	**4.43**	0.01

Na Tabela A.29 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 3 otimiza os índices Dunn, Davies-Bouldin e Silhueta. Todas as soluções obtiveram o mesmo valor par ao índice C.

Tabela A.30: Soluções DLBCL-Outcome pós-MOPSO usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.276	7.30	4.43	0.03
2	0.275	6.99	4.43	0.03
3	0.280	5.60	4.43	0.05
4	0.275	6.63	4.43	0.03
5	0.257	7.49	4.43	0.03
6	0.277	6.03	4.43	0.04
7	0.274	7.30	4.43	0.03

Após o processamento do MOPSO, Tabela A.30, a melhor solução escolhida pelo MOPSO é a de número 3, que também foi selecionada pelos métodos baseado na fórmula de peso e lexicográfico.

A.4. Resultados da Base DLBCL-NIH

Tabela A.31: Quantidade de atributos selecionados para DLBCL-Tumor

Número do método	Quantidade de atributos
1	5
2	36
3	740
4	1850
5	3700
6	5550
7	7399

A.4.1. K-means

Tabela A.32: Soluções DLBCL-NIH usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.41**	0.33
2	0.81	0.36
3	4.00	**0.40**
4	6.03	0.36
5	8.60	0.36
6	10.18	0.36
7	11.19	0.36

Na Tabela A.32 estão as soluções das bases DLBCL-NIH ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento K-*means*. Observa-se que a solução 1 otimizou o critério *Isolation* e a solução 3 otimizou o índice Jaccard.

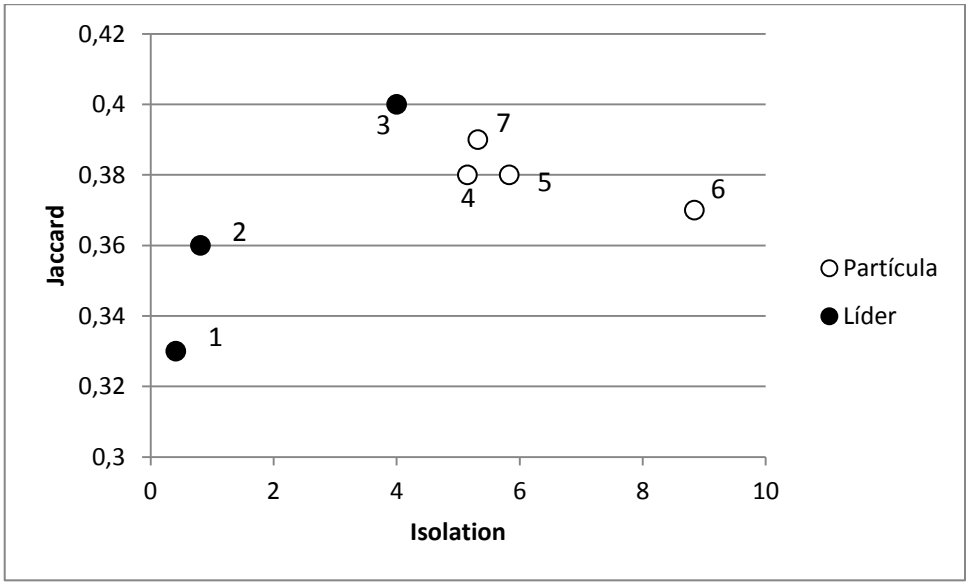

Figura A.10: Soluções DLBCL-NIH pós-MOPSO usando K-means, Isolation e Jaccard

Os líderes escolhidos pela frente de Pareto ao fim do MOPSO, como podem ser vistos na Figura A.10, foram as soluções 1, 2 e 3. A solução 3 dominou todas as partículas não-líderes. Para o método lexicográfico, a melhor solução foi a de número 3, já para o método

baseado na fórmula de peso foi escolhida a solução 2, apesar de não obter melhor valor em nenhum índice.

Tabela A.33: Soluções DLBCL-NIH usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.11	**2.57**	**3.30**	**0.22**
2	0.26	3.34	5.95	0.08
3	0.30	3.72	9.82	0.03
4	0.33	3.73	7.93	0.06
5	0.34	3.53	7.61	0.07
6	0.34	3.57	7.49	0.07
7	**0.35**	3.84	7.68	0.06

Na Tabela A.33 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 7 otimiza, pela primeira vez, o índice Dunn. A solução 1 otimiza os índices Davies-Bouldin, C e Silhueta, tornando a melhor solução para os métodos lexicográfico e baseado na fórmula de peso.

Tabela A.34: Soluções DLBCL-NIH pós-MOPSO usando K-means e critérios realtivos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.11	**2.57**	**3.30**	**0.22**
2	0.26	3.34	5.95	0.08
3	0.31	3.69	9.45	0.04
4	0.34	3.59	7.55	0.07
5	0.34	3.53	7.61	0.07
6	0.34	3.57	7.49	0.07
7	**0.35**	3.84	7.68	0.06

A Tabela A.34 mostra como o MOPSO desloca as soluções. As soluções tidas como ótimas e eleitas líderes pela frente de Pareto foram as de número 1, 2, 5, 6 e 7. Os métodos lexicográfico e baseado em fórmula de peso selecionaram a solução do método C-*FOCUS* como a melhor.

A.4.2. ISODATA

Tabela A.35: Soluções DLBCL-NIH usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**2.23**	**1.0**
2	5.75	0.83
3	25.94	0.33
4	40.25	0.34
5	48.72	0.34
6	65.85	0.25
7	79.99	0.33

Na Tabela A.35 estão as soluções das bases DLBCL-Tumor ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento ISODATA. A solução 1 conseguiu otimizar tanto o índice *Isolation* quanto o índice Jaccard.

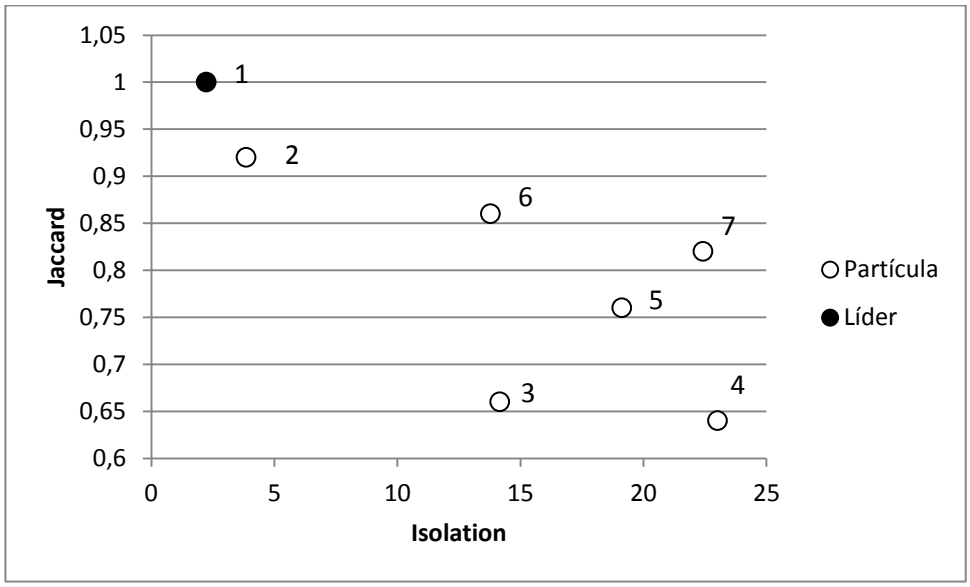

Figura A.11: Soluções DLBCL-NIH pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Figura A.11, a solução 1 domina todas as outras partículas. Por isso continua sendo a única solução ótima para todos os métodos de otimização multiobjetivo.

Tabela A.36: Soluções DLBCL-NIH usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.63**	**1.30**	**3.71**	**0.11**
2	0.33	1.51	6.10	-2.12
3	0.31	1.48	6.86	0.008
4	0.34	1.45	7.26	0.012
5	0.36	1.43	7.05	0.015
6	0.38	1.52	7.16	0.007
7	0.36	1.53	7.10	0.007

Na Tabela A.36 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 1 otimiza todos os índices relativos.

Tabela A.37: Soluções DLBCL-NIH pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**0.63**	**1.30**	**3.71**	**0.11**
2	0.51	1.38	4.63	0.07
3	0.42	1.42	5.82	0.04
4	0.57	1.33	4.40	0.09
5	0.37	1.43	6.96	0.01
6	0.39	1.51	6.94	0.01
7	0.53	1.39	5.02	0.07

Como pode ser observado na Tabela A.37, a solução 1 continua dominando todas as outras soluções, por conseguir otimizar simultaneamente todos os índices.

Com isso, a solução que representa o método de seleção C-*FOCUS*, se tornou a melhor solução para todos os métodos de otimização multiobjetivo.

A.4.3. Classit

Tabela A.38: Soluções DLBCL-NIH usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.0**	**0.52**
2	68.42	0.36
3	67.35	0.35
4	67.59	0.36
5	68.50	0.34
6	68.04	0.35
7	68.48	0.34

Na Tabela A.38 estão as soluções das bases DLBCL-Tumor ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento Classit. A solução 1 conseguiu otimizar tanto o índice *Isolation* quanto o índice Jaccard. Porém, o valor mínimo obtido pelo índice *Isolation* é sinal de que não houve relação entre grupos, pois a soluções 1 gerou apenas um grupo.

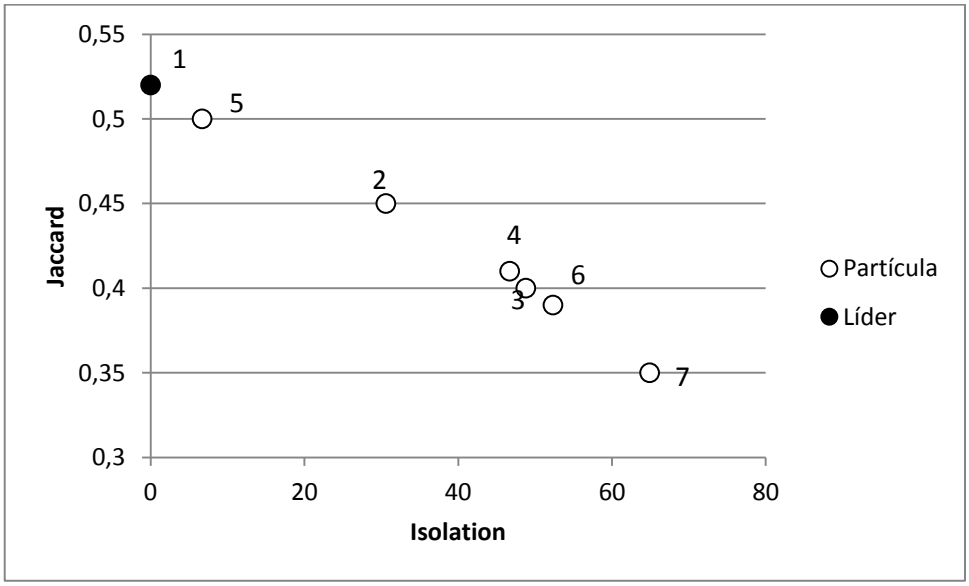

Figura A.12: Soluções DLBCL-NIH pós-MOPSO usando Classit, Isolation e Jaccard

Na Figura A.12, a solução 1 domina todas as outras partículas. Por isso continua sendo a única solução ótima para todos os métodos de otimização multiobjetivo.

Tabela A.39: Soluções DLBCL-NIH usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.79**	**0.0**	**-3.71**	**1.0**
2	0.36	1.54	7.79	0.020
3	0.38	1.56	7.68	0.003
4	0.38	1.55	7.95	0.001
5	0.36	1.55	7.42	-0.003
6	0.37	1.55	7.55	0.003
7	0.36	1.55	7.33	0.002

Na Tabela A.39 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 1 otimiza todos os índices relativos.

Tabela A.40: Soluções DLBCL-NIH pós-MOPSO usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.79**	**0.0**	**-3.71**	**1.0**
2	5.36	1.50	-1.10	0.03
3	1.45	0.29	-3.01	0.81
4	4.82	1.13	-9.97	0.26
5	1.54	0.21	-3.19	0.86
6	1.73	0.05	-3.59	0.96
7	5.93	1.04	-1.22	0.32

Como pode ser observado na Tabela A.40, a solução 1 continua dominando todas as outras soluções, tornando-se na melhor solução para todos os métodos de otimização multiobjetivo.

A.5. Resultados da Base Leukemia-ALL/AML

Tabela A.41: Quantidade de atributos selecionados para Leukemia-ALL/AML

Número do método	Quantidade de atributos
1	1
2	1
3	714
4	1782
5	3565
6	5347
7	7129

A.5.1. K-means

Tabela A.42: Soluções Leukemia-ALL/AML usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.34**	0.5
2	**0.34**	0.5
3	6.54	**0.65**
4	8.82	**0.65**
5	12.7	0.47
6	14.75	0.47
7	15.2	0.45

Na Tabela A.42 estão as soluções das bases Leukemia-ALL/AML ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento K-*means*. Observa-se que as soluções 3 e 4 otimizaram o critério Jaccard e as soluções 1 e 2 otimizaram o índice *Isolation*.

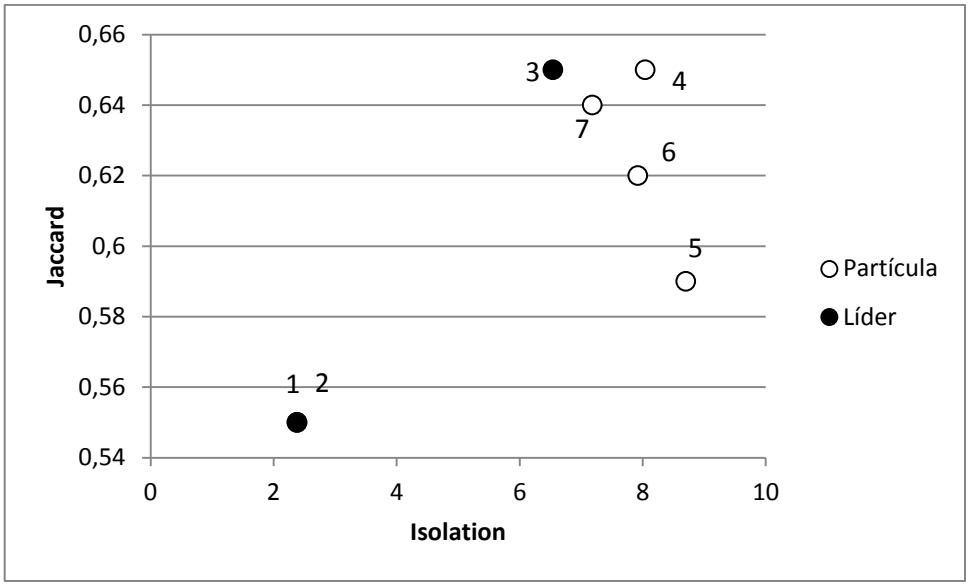

Figura A.13: Soluções Leukemia-ALL/AML pós-MOPSO usando K-means, Isolation e Jaccard

Na Figura A.13 é possível observar como a solução 3 se movimentou e ficou situada na região promissora ao fim da execução do MOPSO. As soluções 1 e 2, por serem soluções iguais, não se dominam mas também não foram dominadas por nenhuma outra solução do enxame, tornando-se também em líderes. Para o método lexicográfico, a melhor solução foi a de número 3, já para o método baseado na fórmula de peso foi escolhida a solução 2. Neste caso, como as soluções 1 e 2 são idênticas, pode-se considerar tanto a solução como a solução 2.

Tabela A.43: Soluções Leukemia-ALL/AML usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.07	**0.57**	**1.0**	**0.61**
2	0.07	**0.57**	**1.0**	**0.61**
3	0.54	1.68	5.04	0.23
4	0.55	2.0	6.06	0.18
5	**0.61**	2.0	10.73	0.07
6	0.59	2.11	10.98	0.06
7	0.56	2.38	9.67	0.07

Na Tabela A.43 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 5 otimiza o índice Dunn, as soluções 1 e 2 otimizam os índices Davies-Bouldin, C e Silhueta.

Tabela A.44: Soluções Leukemia-ALL/AML pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.20	**0.88**	**2.14**	**0.51**
2	0.20	**0.88**	**2.14**	**0.51**
3	0.54	1.68	5.04	0.23
4	0.55	2.00	6.06	0.18
5	**0.61**	2.00	10.73	0.07
6	0.60	2.05	18.85	0.06
7	0.56	2.38	9.67	0.07

A Tabela A.44 mostra como ficaram as posições das soluções após o MOPSO ser executado. As soluções tidas como ótimas e eleitas líderes pela frente de Pareto foram as de número 1, 2, 3, 4, 5 e 6. No método lexicográfico sobraram, por serem iguais, as soluções 1 e 2. O mesmo ocorreu para o método baseado em fórmula de peso.

A.5.2. ISODATA

Tabela A.45: Soluções Leukemia-ALL/AML usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**0.34**	**0.50**
2	**0.34**	**0.50**
3	6.54	0.65
4	8.85	0.60
5	10.66	0.41
6	12.07	0.45
7	12.37	0.35

Na Tabela A.45 estão as soluções das bases Leukemia-ALL/AML ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o ISODATA, otimizadas pelas soluções 1 e 2.

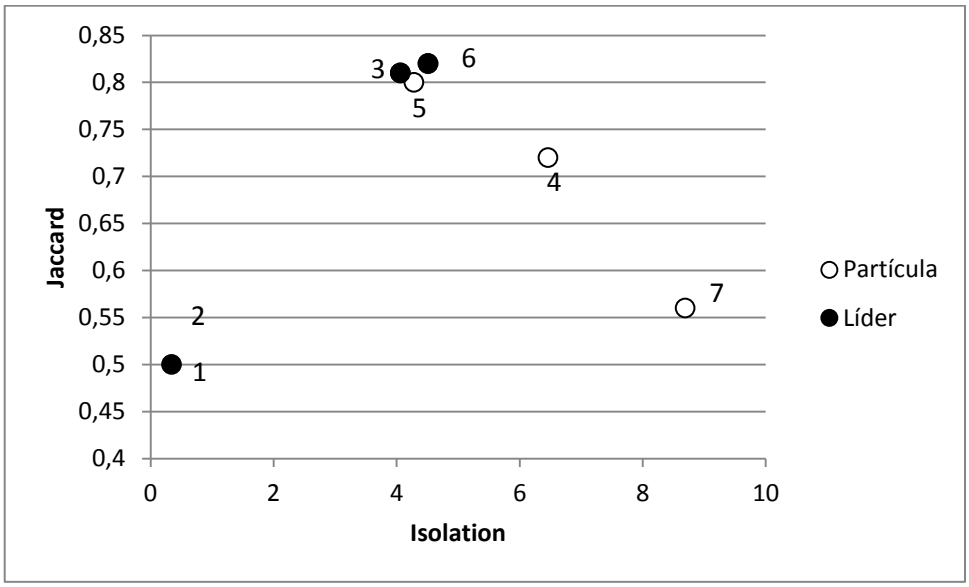

Figura A.14: Soluções Leukemia-ALL/AML pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Figura A.14, as soluções 1, 2, 3 e 6 fazem parte do repositório de líderes ao fim do MOPSO. Tanto o método lexicográfico quanto o baseado na fórmula de peso selecionaram a solução 2 como ótima, também podendo ser considerada a solução 1, por serem idênticas.

Tabela A.46: Soluções Leukemia-ALL/AML usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.14**	**0.15**	**0.36**	**0.93**
2	**1.14**	**0.15**	**0.36**	**0.93**
3	0.54	1.70	5.01	0.23
4	0.48	2.04	5.76	0.20
5	0.51	2.68	7.20	0.10
6	0.51	2.70	7.91	0.08
7	0.42	2.85	7.66	0.10

Tabela A.47: Soluções Leukemia-ALL/AML pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	**1.14**	**0.15**	**0.36**	**0.93**
2	**1.14**	**0.15**	**0.36**	**0.93**
3	1.08	0.31	0.85	0.86
4	0.92	0.80	2.22	0.68
5	1.01	0.69	1.82	0.75
6	0.61	2.31	6.77	0.21
7	0.92	0.98	2.60	0.68

Na Tabela A.46 é possível observar que as soluções 1 e 2 otimizam todos os índices. Assim como na Tabela A.47 e nos demais métodos de otimização multiobjetivo.

A.5.3. Classit

Tabela A.48: Soluções Leukemia-ALL/AML usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	7.28	**0.384**
2	7.28	**0.384**
3	**6.78**	**0.384**
4	7.34	0.360
5	7.47	0.334
6	6.94	0.360
7	7.55	0.334

Na Tabela A.48 estão as soluções das bases Leukemia-ALL/AML ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento Classit. A solução 3 conseguiu otimizar o índice *Isolation*, enquanto as soluções 1, 2 e 3 otimizaram o índice Jaccard.

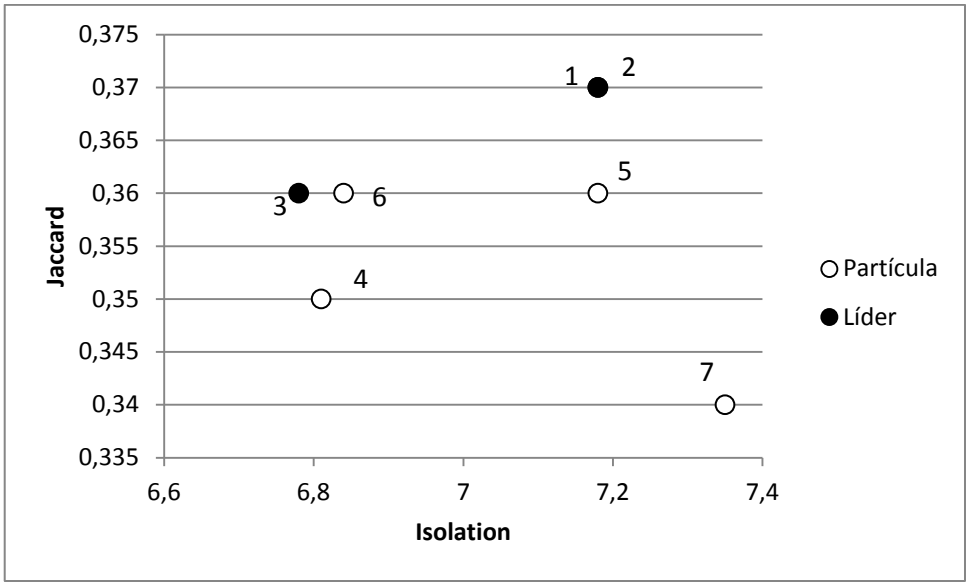

Figura A.15 : Soluções Leukemia-ALL/AML pós-MOPSO usando Classit, Isolation e Jaccard

Na Figura A.15, as soluções 1 e 2 dominam as soluções 5 e 7. Já a solução 3, domina as soluções 6 e 4. Essas duas soluções foram eleitas ótimas pelo MOPSO. Para os métodos lexicográfico e baseado em fórmula de peso, foram selecionadas as soluções 1 e 3, respectivamente.

Tabela A.49: Soluções Leukemia-ALL/AML usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.407	4.90	7.33	0.048
2	0.407	4.90	7.33	0.048
3	**0.429**	5.38	**7.11**	0.046
4	0.382	4.93	**7.11**	0.050
5	0.402	4.86	**7.11**	0.044
6	0.385	5.25	**7.11**	0.044
7	0.398	**4.77**	**7.11**	**0.053**

Na Tabela A.49 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 3 otimiza o índice Dunn, a solução 7 otimiza os índices Davies-Bouldin e Silhueta, e as soluções 3, 4, 5,6 e 7 otimizam o índice C.

Tabela A.50: Soluções Leukemia-ALL/AML pós-MOPSO usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.405	4.88	7.30	0.049
2	0.405	4.88	7.30	0.049
3	**0.429**	5.38	**7.11**	0.046
4	0.401	4.87	**7.11**	0.044
5	0.402	4.86	**7.11**	0.044
6	0.406	5.31	**7.11**	0.045
7	0.398	**4.77**	**7.11**	**0.053**

De acordo com a Tabela A.50, as soluções escolhidas como líderes pelo MOPSO foram as soluções 3, 5 e 7. Para os demais métodos de otimização, foi selecionado a solução 7.

A.6. Resultados da Base Leukemia-MLL

Tabela A.51: Quantidade de atributos selecionados para Leukemia-MLL

Número do método	Quantidade de atributos
1	3
2	113
3	1258
4	3145
5	6291
6	9437
7	12582

A.6.1. K-means

Tabela A.52: Soluções Leukemia-MLL usando K-means, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**2.21**	0.40
2	7.38	**0.59**
3	25.71	0.51
4	48.83	0.53
5	60.36	0.53
6	78.54	0.53
7	86.64	0.53

Na Tabela A.52 estão as soluções das bases Leukemia-MLL ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento K-*means*. Observa-se que a solução 1 otimiza *Isolation*¸ enquanto que a solução 2 otimiza o índice Jaccard.

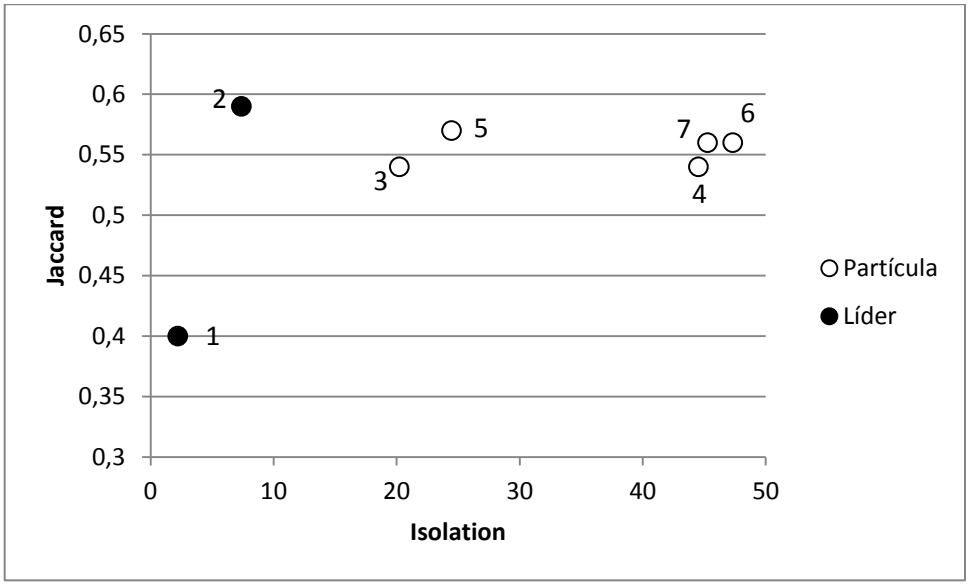

Figura A.16: Soluções Leukemia-MLL pós-MOPSO usando K-means, Isolation e Jaccard

Na Figura A.16 observa-se que não houve mudança de soluções dominantes e as soluções 1 e 2 foram tidas como ótimas. Nos métodos lexicográfico e baseado em fórmula de peso, a melhor solução foi a de número 2.

Tabela A.53: Soluções Leukemia-MLL usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.01	**0.81**	**1.12**	0.11
2	0.44	1.70	5.19	-0.10
3	0.38	1.70	3.36	0.01
4	0.50	1.15	4.64	**0.55**
5	**0.52**	1.55	5.05	-0.12
6	0.51	1.51	5.70	0.46
7	0.51	1.69	5.99	0.43

Observa-se na Tabela A.53 as soluções obtidas pelos critérios relativos, em que a solução 5 otimiza o índice Dunn e Silhueta, e a solução 1 otimizam os índices Davies-Bouldin, e C.

Tabela A.54: Soluções Leukemia-MLL pós-MOPSO usando K-means e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.01	**0.81**	**1.12**	0.11
2	0.45	1.68	5.17	-0.11
3	0.38	1.70	3.36	0.01
4	0.50	1.15	4.64	**0.55**
5	**0.52**	1.55	5.05	-0.12
6	0.51	1.51	5.70	0.46
7	0.51	1.65	5.91	0.44

A Tabela A.54 mostra como ficaram as posições das soluções após o MOPSO ser executado. As únicas soluções a não serem consideradas ótimas pelo MOPSO foram a de número 7, que representa a base original, e a base 2, representante da base reduzida pelo CFS.

O método lexicográfico elegeu a solução 1 como melhor, enquanto que o método baseado em fórmula de peso selecionou a solução 4.

A.6.2. ISODATA

Tabela A.55: Soluções Leukemia-MLL usando ISODATA, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	**1268**	0.33
2	25963	0.42
3	94980	**0.48**
4	116875	0.40
5	272852	0.22
6	150156	0.32
7	156634	0.32

Na Tabela A.55 estão as soluções das bases Leukemia-MLL ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento ISODATA. A solução 1 conseguiu otimizar o índice *Isolation*, enquanto a solução 3 otimizou o índice Jaccard.

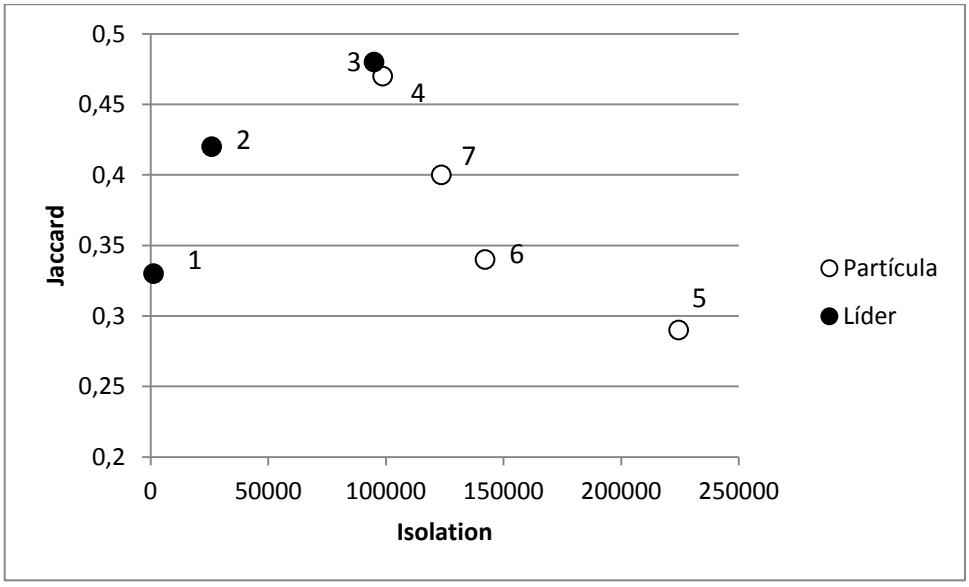

Figura A.17: Soluções Leukemia-MLL pós-MOPSO usando ISODATA, Isolation e Jaccard

Na Figura A.17, a solução 1 domina a solução 5, que foi classificada como a pior. A solução 2 dominou as soluções 7 e 6, já a solução 3 dominou a solução 4. Para o método lexicográfico, a melhor solução foi a de número 3, enquanto que para o método baseado em fórmula de peso foi a de número 2.

Tabela A.56: Soluções Leukemia-MLL usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.03	**0.99**	**2.72**	**8.13**
2	0.32	1.00	3.59	-0.33
3	0.31	**0.99**	3.84	6.42
4	0.27	**0.99**	4.37	5.84
5	0.31	1.00	3.54	-0.33
6	0.32	**0.99**	5.34	5.93
7	**0.33**	1.00	4.27	-0.33

Na Tabela A.56 é possível observar as soluções obtidas pelos critérios relativos, em que a solução 7 otimiza o índice Dunn, as soluções 1,3,4 e 6 otimizaram o índice Davies-Bouldin, a solução 1 otimiza o índice C, além do índice Silhueta.

Tabela A.57: Soluções Leukemia-MLL pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.03	**0.99**	**2.70**	**8.13**
2	0.32	1.00	3.59	-0.33
3	0.31	**0.99**	3.84	6.42
4	0.30	1.00	4.32	-0.16
5	0.31	1.00	3.54	-0.33
6	0.32	**0.99**	5.34	5.93
7	**0.33**	1.00	4.27	-0.33

Como pode ser observado na Tabela A.57, a única solução que não se encontra na região promissora do enxame é a de número 4. A solução obtida pela base original (solução 7) foi eleita a melhor pelo método baseado em fórmula de peso e, para o método lexicográfico, a melhor solução é a de número 1.

A.6.3. Classit

Tabela A.58: Soluções Leukemia-MLL usando Classit, Isolation e Jaccard

Número do método	Isolation	Jaccard
1	22.32	0.18
2	24.58	0.19
3	**20.16**	0.19
4	21.66	0.21
5	29.74	**0.22**
6	23.14	0.20
7	23.69	0.18

Na Tabela A.58 estão as soluções das bases Leukemia-MLL ao serem avaliadas pelos critérios *Isolation* e Jaccard usando o método de agrupamento Classit. A solução 3 conseguiu otimizar o índice *Isolation*, enquanto a solução 5 otimizou o índice Jaccard.

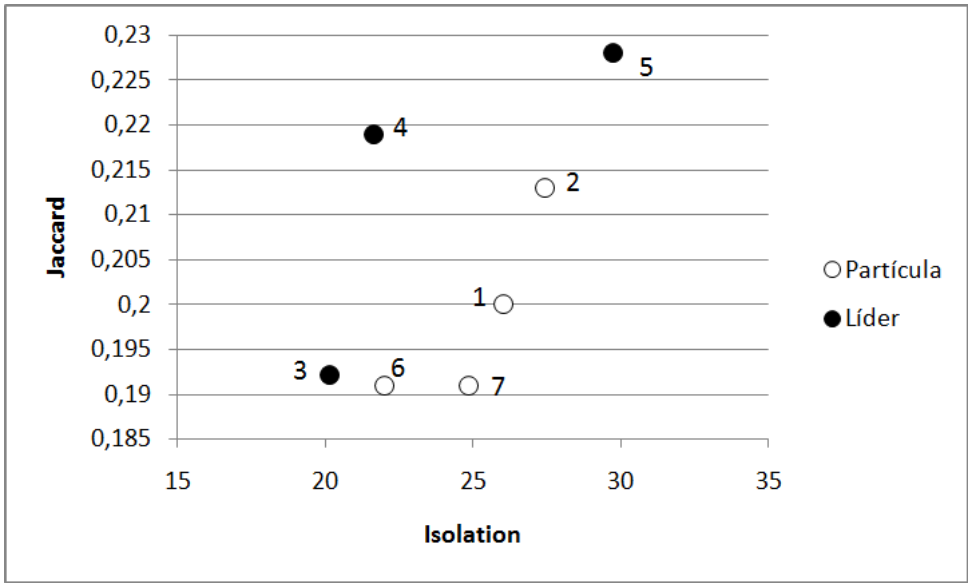

Figura A.18: Soluções Leukemia-MLL pós-MOPSO usando Classit, Isolation e Jaccard

Na Figura A.18, a solução 4 domina as demais soluções, exceto as soluções 5 e 3, que não são dominadas por nenhuma partícula do enxame. Ao final do MOPSO, a frente de Pareto selecionou as soluções 3, 4 e 5 como líderes. O método lexicográfico selecionou a solução 5, enquanto que o método baseado em fórmula de peso selecionou a solução 4.

Tabela A.59: Soluções Leukemia-MLL usando Classit e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.344	6.06	4.03	-0.015
2	0.335	6.00	**4.01**	-0.005
3	0.335	7.36	4.03	0.006
4	0.331	6.72	4.03	-0.013
5	**0.348**	**4.57**	4.03	**0.014**
6	0.335	5.90	4.03	-0.009
7	0.335	6.00	4.03	-0.012

Na Tabela A.59 é possível observar as soluções obtidas pelos critérios relativos, em que a solução a solução 5 otimiza os índices Dunn, Davies-Bouldin e Silhueta, e a solução 2 otimiza o índice C.

Tabela A.60: Soluções Leukemia-MLL pós-MOPSO usando ISODATA e critérios relativos

Número do método	Dunn	Davies-Bouldin	C	Silhueta
1	0.336	7.13	**4.01**	0.002
2	0.335	7.30	**4.01**	0.005
3	0.335	7.36	**4.01**	0.006
4	0.334	7.16	**4.01**	-1.80
5	**0.348**	**4.57**	4.03	**0.014**
6	0.335	7.19	**4.01**	0.004
7	0.338	7.02	**4.01**	0.001

Como pode ser observado na Tabela A.60, a solução 5 otimiza os índices Dunn, Davies-Bouldin e Silhueta. Por isso, essa solução, juntamente com a solução 3, compõem o conjunto de líderes do MOPSO. Tanto o método lexicográfico, quanto o método baseado em fórmula de peso selecionaram a solução 5 como ótima.

Apêndice B

Genes Mais Selecionados

Nesta parte do trabalho serão apresentados os genes que foram selecionados por diversas vezes. Eles estão nas tabelas a seguir, que são compostas pelo nome do gene (de acordo com as descrições das bases) e em quais métodos de seleção de atributos eles aparecem.

Com as informações fornecidas nas próximas tabelas, especialistas nas doenças abordadas no trabalho (linfoma e leucemia) podem analisar se os genes tidos como relevantes nos experimentos são realmente importantes para a aparição ou inibição dessas doenças.

Como foram usadas quatro variações do método de seleção *Relief*-F, todos os genes selecionados pelo *Relief*-F utilizando 10% da quantidade total de atributos, por exemplo, estarão nas demais seleções realizadas por esse método.

Os genes que serão apresentados a seguir foram selecionados por pelo menos dois métodos de seleção diferentes, considerando apenas o C-*FOCUS*, o CFS e a variação *Relief*-F 10% para manter o maior nível de relevância possível.

B.1. DLBCL-Stanford

Na Tabela B.1 estão os genes mais selecionados para a base DLBCL-Stanford.

Tabela B.1: Genes mais escolhidos da base DLBCL-Stanford

Nome do Gene	Métodos que selecionaram o gene
GENE2544X	C-*FOCUS* e *Relief*-F 10%
GENE3327X	C-*FOCUS* e *Relief*-F 10%

GENE2513X	CFS e *Relief*-F 10%
GENE2513X	CFS e *Relief*-F 10%
GENE2294X	CFS e *Relief*-F 10%
GENE2322X	CFS e *Relief*-F 10%
GENE2329X	CFS e *Relief*-F 10%
GENE2065X	CFS e *Relief*-F 10%
GENE3324X	CFS e *Relief*-F 10%
GENE3325X	CFS e *Relief*-F 10%
GENE3330X	CFS e *Relief*-F 10%
GENE3332X	CFS e *Relief*-F 10%
GENE3259X	CFS e *Relief*-F 10%
GENE3256X	CFS e *Relief*-F 10%
GENE3261X	CFS e *Relief*-F 10%
GENE1211X	CFS e *Relief*-F 10%
GENE1212X	CFS e *Relief*-F 10%
GENE446X	CFS e *Relief*-F 10%
GENE427X	CFS e *Relief*-F 10%
GENE404X	CFS e *Relief*-F 10%
GENE86X	CFS e *Relief*-F 10%
GENE67X	CFS e *Relief*-F 10%
GENE3821X	CFS e *Relief*-F 10%
GENE1719X	CFS e *Relief*-F 10%
GENE1720X	CFS e *Relief*-F 10%
GENE1567X	CFS e *Relief*-F 10%
GENE3933X	CFS e *Relief*-F 10%

Apesar da grande quantidade de atributos selecionados mais de uma vez, não houve nenhum que tenha sido selecionado pelos três métodos de seleção.

B.2. DLBCL-Tumor

Na Tabela B.2 estão os genes mais selecionados para a base DLBCL-Tumor.

Tabela B.2: Genes mais escolhidos da base DLBCL-Tumor

Nome do Gene	Métodos que selecionaram o gene
AC002073_cds1_at	C-*FOCUS* e *Relief*-F 10%
D55716_at	C-*FOCUS*, CFS e *Relief*-F 10%
M10901_at	C-*FOCUS* e *Relief*-F 10%
AF008937_at	CFS e *Relief*-F 10%
D25328_at	CFS e *Relief*-F 10%
D64154_at	CFS e *Relief*-F 10%
D79997_at	CFS e *Relief*-F 10%
D83597_at	CFS e *Relief*-F 10%
D87119_at	CFS e *Relief*-F 10%
D87445_at	CFS e *Relief*-F 10%
J03909_at	CFS e *Relief*-F 10%
L06419_at	CFS e *Relief*-F 10%
L22343_at	CFS e *Relief*-F 10%
L25876_at	CFS e *Relief*-F 10%
L27071_at	CFS e *Relief*-F 10%
L42324_at	CFS e *Relief*-F 10%
M12759_at	CFS e *Relief*-F 10%
M15205_at	CFS e *Relief*-F 10%
M31520_at	CFS e *Relief*-F 10%
M35878_at	CFS e *Relief*-F 10%
M63379_at	CFS e *Relief*-F 10%
M65290_at	CFS e *Relief*-F 10%
U50535_at	CFS e *Relief*-F 10%
U61262_at	CFS e *Relief*-F 10%
U64863_at	CFS e *Relief*-F 10%

U81375_at	CFS e *Relief*-F 10%
X01060_at	CFS e *Relief*-F 10%
X02152_at	CFS e *Relief*-F 10%
X03066_at	CFS e *Relief*-F 10%
X16396_at	CFS e *Relief*-F 10%
X16983_at	CFS e *Relief*-F 10%
X69433_at	CFS e *Relief*-F 10%
X82240_rna1_at	CFS e *Relief*-F 10%
X85785_rna1_at	CFS e *Relief*-F 10%
Z21966_at	CFS e *Relief*-F 10%
L33930_s_at	CFS e *Relief*-F 10%
V00594_s_at	CFS e *Relief*-F 10%
M26004_s_at	CFS e *Relief*-F 10%
M88461_s_at	CFS e *Relief*-F 10%
M16652_at	CFS e *Relief*-F 10%
X81836_s_at	CFS e *Relief*-F 10%
M94880_f_at	CFS e *Relief*-F 10%

Como pôde ser observado, o atributo D55716_at foi selecionado em todos os métodos de seleção de atributo.

B.3. DLBCL-Outcome

Na Tabela B.3 estão os genes mais selecionados para a base DLBCL-Outcome.

Tabela B.3: Genes mais escolhidos da base DLBCL-Outcome

Nome do Gene	Métodos que selecionaram o gene
AFFX-BioC-3_at	C-*FOCUS,* CFS e *Relief*-F 10%
M83186_at	C-*FOCUS* e CFS
U23028_at	C-*FOCUS* e CFS
U83908_at	C-*FOCUS,* CFS e *Relief*-F 10%

X77307_at	C-*FOCUS,* CFS e *Relief*-F 10%
U48865_s_at	C-*FOCUS,* CFS e *Relief*-F 10%
AC002450_at	CFS e *Relief*-F 10%
D80003_at	CFS e *Relief*-F 10%
L40377_at	CFS e *Relief*-F 10%
M83941_at	CFS e *Relief*-F 10%
U36621_cds2_at	CFS e *Relief*-F 10%
U51903_at	CFS e *Relief*-F 10%
U66702_at	CFS e *Relief*-F 10%
U70663_at	CFS e *Relief*-F 10%
U77413_at	CFS e *Relief*-F 10%
U88871_at	CFS e *Relief*-F 10%
X14046_at	CFS e *Relief*-F 10%
X83412_at	CFS e *Relief*-F 10%
Z48481_at	CFS e *Relief*-F 10%
Z49995_at	CFS e *Relief*-F 10%
X55005_rna1_at	CFS e *Relief*-F 10%
X70811_at	CFS e *Relief*-F 10%
HG4020-HT4290_s_at	CFS e *Relief*-F 10%
U12259_cds2_s_at	CFS e *Relief*-F 10%
U33936_s_at	CFS e *Relief*-F 10%
U90543_at	CFS e *Relief*-F 10%
M99435_at	CFS e *Relief*-F 10%
U46744_at	CFS e *Relief*-F 10%

Observa-se que os atributos AFFX-BioC-3_at, U83908_at, X77307_at e U48865_s_at fazem parte do conjunto selecionado por todos os métodos de seleção.

B.4. DLBCL-NIH

Na Tabela B.4 estão os genes mais selecionados para a base DLBCL-NIH.

Tabela B.4: Genes mais escolhidos da base DLBCL-NIH

Nome do Gene	Métodos que selecionaram o gene
30226	C-*FOCUS,* CFS e *Relief*-F 10%
15875	C-*FOCUS,* CFS e *Relief*-F 10%
26458	C-*FOCUS* e CFS
17385	C-*FOCUS* e CFS
30624	CFS e *Relief*-F 10%
33707	CFS e *Relief*-F 10%
30385	CFS e *Relief*-F 10%
32121	CFS e *Relief*-F 10%
31952	CFS e *Relief*-F 10%
17488	CFS e *Relief*-F 10%
31473	CFS e *Relief*-F 10%
17950	CFS e *Relief*-F 10%
27444	CFS e *Relief*-F 10%
26207	CFS e *Relief*-F 10%
27635	CFS e *Relief*-F 10%
24787	CFS e *Relief*-F 10%
29194	CFS e *Relief*-F 10%
26714	CFS e *Relief*-F 10%
28812	CFS e *Relief*-F 10%
27161	CFS e *Relief*-F 10%

Na base DLBCL-NIH, os atributos 30226 e 15875 foram selecionados por todos os métodos de seleção.

B.5. Leukemia-AML/ALL

Na Tabela B.5 estão os genes mais selecionados para a base Leukemia-AML/ALL. Nela pode-se observar que o mesmo atributo foi selecionado pelos métodos C-*FOCUS* e CFS. Este também faz parte do conjunto de atributos selecionados pela *Relief*-F 10%.

Tabela B.5: Genes mais escolhidos da base Leukemia-AML/ALL

Nome do Gene	Métodos que selecionaram o gene
attribute6855	C-*FOCUS,* CFS e *Relief*-F 10%

B.6. Leukemia-MLL

Na Tabela B.6 estão os genes mais selecionados para a base Leukemia-MLL. Apesar da grande quantidade de atributos que fazem parte do conjunto selecionado por pelo menos dois métodos de seleção, apenas o atributos 36239_at está presente na seleção de todos os métodos.

Tabela B.6: Genes mais escolhidos da base Leukemia-MLL

Nome do Gene	Métodos que selecionaram o gene
35505_at	C-*FOCUS* e *Relief*-F 10%
36239_at	C-*FOCUS,* CFS e *Relief*-F 10%
31481_s_at	CFS e *Relief*-F 10%
31521_f_at	CFS e *Relief*-F 10%
31575_f_at	CFS e *Relief*-F 10%
31736_at	CFS e *Relief*-F 10%
34168_at	CFS e *Relief*-F 10%
34584_at	CFS e *Relief*-F 10%
36386_at	CFS e *Relief*-F 10%
34046_at	CFS e *Relief*-F 10%
34912_at	CFS e *Relief*-F 10%
35926_s_at	CFS e *Relief*-F 10%
36777_at	CFS e *Relief*-F 10%

38518_at	CFS e *Relief*-F 10%
38604_at	CFS e *Relief*-F 10%
40313_at	CFS e *Relief*-F 10%
41078_at	CFS e *Relief*-F 10%
41110_at	CFS e *Relief*-F 10%
41448_at	CFS e *Relief*-F 10%
41624_r_at	CFS e *Relief*-F 10%
41853_at	CFS e *Relief*-F 10%
31816_at	CFS e *Relief*-F 10%
34699_at	CFS e *Relief*-F 10%
35165_at	CFS e *Relief*-F 10%
35985_at	CFS e *Relief*-F 10%
36553_at	CFS e *Relief*-F 10%
36878_f_at	CFS e *Relief*-F 10%
36897_at	CFS e *Relief*-F 10%
37539_at	CFS e *Relief*-F 10%
39749_at	CFS e *Relief*-F 10%
40493_at	CFS e *Relief*-F 10%
40782_at	CFS e *Relief*-F 10%
34322_r_at	CFS e *Relief*-F 10%
34833_at	CFS e *Relief*-F 10%
35261_at	CFS e *Relief*-F 10%
36137_at	CFS e *Relief*-F 10%
36209_at	CFS e *Relief*-F 10%
36678_at	CFS e *Relief*-F 10%
37027_at	CFS e *Relief*-F 10%
37043_at	CFS e *Relief*-F 10%
37332_r_at	CFS e *Relief*-F 10%
37376_at	CFS e *Relief*-F 10%

39556_at	CFS e *Relief*-F 10%
41346_at	CFS e *Relief*-F 10%
41489_at	CFS e *Relief*-F 10%
1389_at	CFS e *Relief*-F 10%
1307_at	CFS e *Relief*-F 10%
873_at	CFS e *Relief*-F 10%
317_at	CFS e *Relief*-F 10%